Tesis de Maestría

Aplicaciones de data mining al estudio de la biodiversidad en relevamientos metagenómicos

Santa María, Cristóbal Raúl

2011

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Aplicaciones de Data Mining al Estudio de la Biodiversidad en Relevamientos Metagenómicos

Tesis presentada para obtener el título de
Magister en Explotación de Datos y Descubrimiento del Conocimiento

Cristóbal Raúl Santa María
Departamento de Ingeniería e Investigaciones Tecnológicas. UNLAM
Cátedra de Microbiología Agrícola. Facultad de Agronomía. UBA

Director de Tesis: Dr. Marcelo A. Soria
Cátedra de Microbiología Agrícola. Facultad de Agronomía. UBA
Maestría en Explotación de Datos y Descubrimiento del Conocimiento
Facultad de Ciencias Exactas y Naturales. UBA

Buenos Aires, 2011

A Ana, Victoria, Javier y Facundo

El trabajo desarrollado para esta Tesis fue realizado con el aporte económico del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza

"Cuentan con los dedos uno, dos, tres, cuatro, muchos; el infinito empieza en el pulgar"
del Informe de Brodie. J. L. Borges

RESUMEN

El trabajo aquí presentado trata acerca de las mediciones de biodiversidad en comunidades microbianas que suelen involucrar dos aspectos: la riqueza y la distribución de los taxones. Una metodología usual para estudiar esas comunidades comprende la utilización de genes marcadores, tal como el que codifica para el rRNA 16S. Se presenta un estado del arte referido a las técnicas de procesamiento computacional que son empleadas, en esos análisis, sobre las secuencias de ADN del gen marcador. También se reseñan las formas de estimación estadística de la diversidad más comúnmente usadas. Se evalúan y detallan las limitaciones que surgen de la aplicación de esos métodos, que comprenden procedimientos habituales en explotación de datos afectados, en este caso, por la presencia de taxones dominantes y de otros que resultan raros aunque no menos importantes desde el punto de vista del análisis del ecosistema. Se proponen alternativas de estimación por simulación para el descubrimiento del conocimiento sobre cantidad de taxones y distribución de los mismos. Los estimadores desarrollados procuran describir las características de la comunidad hallando un patrón distintivo a partir de los datos. En particular se utiliza una idea de Alan Turing acerca de la probabilidad de selección de una especie aun no contabilizada, para construir un Algoritmo de Recuento de Especies (ARE) que expande la muestra original poniendo en evidencia la distribución real y la riqueza. Se emplea también la idea de cobertura muestral para proponer distintas correcciones a este procedimiento y se construye un algoritmo de estimación que combina el uso de ambos estimadores con el de la entropía, que mide la cantidad de información muestral. Los resultados de las pruebas realizadas muestran el desempeño más eficiente de los algoritmos construidos respecto de las mediciones por estimación no paramétrica o por rarefacción, las que a menudo subestiman los valores de riqueza de la población microbiana.

ÍNDICE

1- PRESENTACIÓN…….. 5

2- INTRODUCCIÓN
2.1 ADN…….. 7
2.2 METAGENÓMICA…….. 7
2.3 FILOGENÉTICA……..… 9
2.4 BIODIVERSIDAD…….....12

3- ESTADO DEL ARTE
3.1 EL MODELO MULTINOMIAL Y LA ESTIMACIÓN DE LA RIQUEZA…. 16
3.2 ESTIMACIONES NO PARAMÉTRICAS……..19
3.3 CURVAS DE RAREFACCIÓN……..20
3.4 PROCESOS PREVIOS A LA ESTIMACIÓN……..….23
3.5 ANÁLISIS DEL PROBLEMA……....28
3.6 CONCLUSIONES…….....30

4- LOS DATOS Y EL PROCESO ESTÁNDAR
4.1 CONJUNTOS DE MUESTRAS……...31
4.2 PROCESOS INICIALES……..32
4.3 EVALUACION DE RIQUEZA Y DIVERSIDAD……..35

5- SIMULACIÓN
5.1 EL MODELO EXPERIMENTAL……39
5.2 ALGORITMO DE RECUENTO DE ESPECIES (ARE)……..41
5.3 ALGORITMO DE RECUENTO DE ESPECIES CON COBERTURA (AREC)……..……43
5.4 SUAVIZACIONES (ARECS1 Y ARECS2) ……..…….47
5.5 PRUEBAS…….49
5.6 USO DEL COEFICIENTE DE VARIACIÓN (ARECV)…….....57
5.7 USO DE LA ENTROPÍA (AREN Y ARECE)……...……..61

6- CONCLUSIONES
6.1 CONSIDERACIONES GENERALES…….….67
6.2 PERSPECTIVAS……...74

BIBLOGRAFIA……...……...75

ANEXO
A- PROGRAMAS…….79
B- ARTÍCULOS PRESENTADOS A CONGRESOS……..80

1. PRESENTACIÓN

En los estudios microbiológicos de comunidades por lo general es importante conocer la cantidad de especies presentes en el medio y su distribución. En otras oportunidades se trata de saber cuantos grupos taxonómicos de microorganismos están representados en la comunidad y cual es la proporción en que esto ocurre para cada caso. En forma global, puede decirse que, para evaluar la biodiversidad de un medio, hay que calcular las cantidades de taxones, ya sean especies, géneros, familias u otros, a efecto de establecer su riqueza y analizar además la forma en se distribuyen.

La tarea requiere acordar, en primer lugar, un criterio biológico para identificar los taxones. Una alternativa crecientemente utilizada al respecto es el análisis basado en el gen 16S rRNA que ha tenido una alta conservación a lo largo del proceso evolutivo y que permite, por ello, apreciar con exactitud las diferencias taxonómicas [1]. Una vez secuenciadas las cadenas de ADN del gen, obtenidas de una muestra de material biológico, estas pueden alinearse de acuerdo a distintos patrones. Luego pueden medirse las "distancias genéticas" entre secuencias para realizar un agrupamiento en "clusters" según el grado de similaridad que revelen. Los distintos umbrales de disimilaridad que se eligen para formar estos grupos permiten establecer el nivel taxonómico, especie o familia por ejemplo, al cual se realiza el estudio. Es decir, finalmente para contar cantidad de taxones y averiguar su distribución en la muestra de material tomada, habrá que contar "clusters" y cantidad de secuencias que conforman a cada uno de ellos.

Pero, cuando se desea inferir desde una muestra la riqueza de todo el medio biológico, se presentan además otras dificultades, de carácter estadístico, que provienen de la gran cantidad de microorganismos que integran realmente la comunidad y de la existencia de taxones que se encuentran en muy baja proporción y resultan, por ende, raros. Ocurre entonces que el tamaño de la población y la rareza estadística de algunos taxones, cuya importancia biológica puede ser mucho más significativa que su número, se suman a limitaciones tecnológicas y/o económicas para introducir un grado de incertidumbre en las estimaciones de biodiversidad poblacional a partir de muestras, que no puede tratarse con las técnicas estadísticas habituales.

Existen distintos modelos a partir de los cuales es posible abordar las situaciones planteadas, pero sus resultados suelen subestimar la real cantidad de taxones presentes en la comunidad y, por lo tanto, desconocer una parte de la distribución de los

mismos [2]. De acuerdo a ello, la idea del presente trabajo es explorar algunas alternativas, desde la perspectiva de los datos existentes en la muestra inicial, que permitan estimar la riqueza y distribución de taxones, dos de los parámetros mas importantes al caracterizar una comunidad microbiana, aportando mayor precisión en sus determinaciones por vía de la aplicación de técnicas de explotación de datos y simulación estadística combinadas.

Finalizada esta presentación, la exposición del trabajo realizado continúa con un segundo capítulo que introduce el contexto en el que se desarrolló la investigación en cuanto a teorías y técnicas sobre ADN, metagenómica, filogenética y biodiversidad. En el tercer capítulo se analiza el "estado del arte" respecto de las mediciones de biodiversidad considerándose el modelo empleado habitualmente, las estimaciones no paramétricas existentes y la técnica de estimación por curvas de rarefacción. Se analiza también con detalle el modelo filogenético utilizado para calcular las distancias y realizar agrupamientos de casos por taxones. Los últimos apartados del capítulo están dedicados a comentar los problemas y limitaciones que plantean las técnicas expuestas para medir la biodiversidad. El cuarto capítulo presenta los conjuntos de datos con los cuales se trabajará procurando mejorar las estimaciones de riqueza y distribución biológicas. También detalla los procesos que se realizan sobre las muestras, previos a las estimaciones. Luego, para los conjuntos seleccionados, se calculan los valores de los estimadores de riqueza y diversidad en general, más comúnmente empleados y se analizan los resultados. En el quinto capítulo se introduce el modelo experimental que habrá de usarse para explotar los datos y descubrir los patrones de riqueza y distribución en comunidades microbianas. Se construyen siete algoritmos alternativos para estimar la riqueza y la diversidad en general. En cada caso se analizan los resultados obtenidos al ser aplicados sobre los conjuntos de muestras seleccionados. El sexto y último capítulo está dedicado a exponer las conclusiones del trabajo. Finalmente se presenta la bibliografía referida en la exposición y otra utilizada para la comprensión de la temática abordada. Se incorpora además un CD anexo con los distintos programas de computadora, realizados en lenguaje R, que contiene también los trabajos presentados en los Workshops de Investigadores en Ciencias de la Computación (WICC 2010 y 2011) y en el XVII Congreso Argentino de Ciencias de la Computación (CACIC 2011).

2. INTRODUCCIÓN

2.1 ÁCIDO DESOXIRRIBONUCLEICO (ADN)

En forma esquemática puede decirse que la biología molecular investiga la estructura y función de las proteínas y los ácidos nucleicos en los organismos vivos. Esto incluye necesariamente el estudio de los genes de cada organismo que se encuentran presentes en los cromosomas dentro de las células.

El ADN (ácido desoxirribonucleico) es una molécula lineal extremadamente larga, un polímero, compuesto por la sucesión de cuatro componentes modulares o monómeros: los desoxinucleótidos, que pueden ser adenina (A), citosina (C), guanina (G) o timina (T). En la mayoría de los seres vivos funciona como repositorio de la información genética codificada en forma de "mensajes" constituidos por secuencias de los cuatro componentes mencionados. Esta información es necesaria para que la célula sintetice las proteínas que, a su vez, dan forma a los organismos o funcionan como catalizadores de reacciones metabólicas. Ver Durbin et al. [3]

El ADN total de un procariota, también llamado genoma, puede abarcar unos cuatro millones de nucleótidos, en una cadena circular. Al secuenciar un genoma los métodos de laboratorio en uso permiten obtener cadenas mucho más cortas, de entre 50 y 800 nucleótidos, según la tecnología empleada. Para secuenciar el genoma completo de un organismo se fragmenta el ADN en el laboratorio y se obtienen cientos de miles, a veces millones, de segmentos al azar con superposición parcial. Una vez secuenciados, los fragmentos se ensamblan aplicando alguno de los varios métodos computacionales de ensamblado. También existen metodologías de laboratorio para secuenciar y analizar genes individuales cuyo uso se comenta más adelante.

2.2 METAGENOMICA

La metagenómica, cuyo desarrollo comienza con el actual siglo, realiza el análisis genómico de comunidades microbianas. Combina el concepto estadístico de meta-análisis referido al proceso en el que se relacionan estadísticamente análisis separados, con la genómica que es el análisis comprensivo del material genético de un organismo. Este nuevo campo trata de explorar un conjunto de secuencias de ADN que

constituyen parte del material genético de una comunidad y que responden a genomas de distintos organismos. Ver Liza Gross [4].

Si se tiene en cuenta que el número estimado de procariotas (organismos unicelulares sin núcleo) presentes en el planeta es mayor que 10^{29} se comprende que, en los distintos ecosistemas biológicos, se hallen en cantidades considerables. Guazzaroni et al. [5] Todos poseen una estructura genética y algunos tienen especial incidencia en los procesos de transformación química que ocurren en el medio que habitan aunque, actualmente, la mayoría de estos microorganismos no pueden ser cultivados en laboratorio a efecto de extraer su ADN e investigar su genoma. Se estima que sólo un 1% de todas las especies bacterianas que habitan el planeta se pueden cultivar y estudiar aisladas en el laboratorio. Sin embargo, es posible determinar su existencia a través de su ADN. Las técnicas descriptas para obtener ADN a partir de muestras ambientales no requieren cultivo y constituyen hoy en día la mejor herramienta para establecer la existencia de bacterias y los ambientes en que se desarrollan. Schloss y Handelsman [6].

Los análisis metagenómicos se pueden dividir en dos grandes categorías. La primera incluye a aquellos estudios por los cuales se intenta obtener la mayor cantidad de secuencias posibles, operando en una forma análoga al secuenciado de un único genoma descripto anteriormente. Así, la muestra contiene fragmentos provenientes de muchos genomas diferentes. En general, dada una muestra de una comunidad, no es posible ensamblar genomas individuales, pero se obtienen conjuntos de datos con secuencias parciales que en algunos casos se pueden asignar a especies conocidas. Esta estrategia se denomina con sus siglas WGS, del inglés “Whole Genome Sequencing”. El segundo tipo de análisis metagenómico se conoce con el nombre de “análisis de marcadores”, y es el que se empleará en este trabajo. El objetivo de este tipo de estudio es obtener secuencias de uno, o unos pocos, genes predeterminados, para establecer, cuando sea posible, a qué especie pertenecen.

Los análisis de marcadores son además, un buen punto de partida para realizar estudios comparativos de biodiversidad Esto es así porque la distinción entre especies resulta claramente señalada por disimilaridades aparecidas en secuencias correspondientes al gen marcador en particular. Estas secuencias por lo general, han sido bien conservadas en el desarrollo evolutivo y presentan mínimas variaciones que indican la diferencia de especies. Tal es el caso del gen que codifica para el ARN ribosomal de 16S (16S rRNA), el más común de los marcadores presente en procariotas, cuyo uso en estimaciones comparativas de riqueza puede verse por ejemplo en los

trabajos de Youssef y Elshahed [1] y White et al. [7] Además este gen es utilizado en estudios llamados filogenéticos que tratan de estimar la historia evolutiva del desarrollo de las especies. Se comparan secuencias y se ven las diferencias para estructurar árboles filogenéticos y secuencias evolutivas al partir de una división en ramas denominadas bacterias y arqueobacterias, estas dos últimas procariotas. Ver Brady y Salzberg [8]

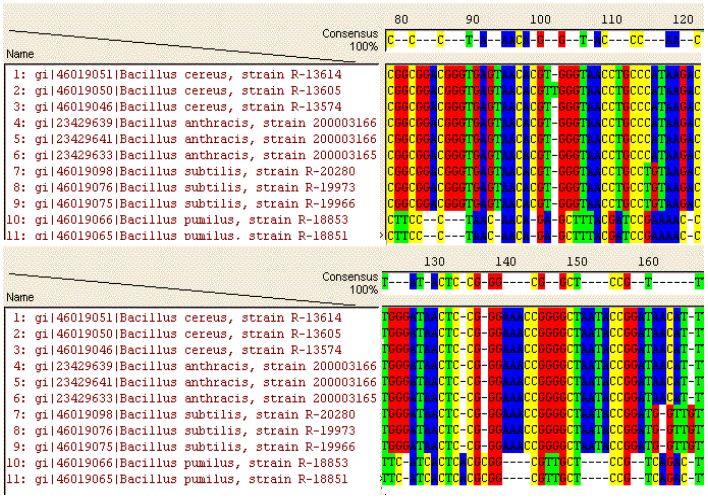
2.3 FILOGENETICA

Ya se ha mencionado que el gen 16S rRNA se utiliza como marcador pues su estructura general se ha conservado a través de la historia evolutiva. Está presente en bacterias y arqueobacterias, y su secuencia sufre transformaciones que responden a procesos de cambio evolutivo. Se asume que aquellos individuos con secuencias más similares entre sí están emparentados evolutivamente y comparten un ancestro más cercano que dos individuos con menor similitud entre sus secuencias. Las relaciones evolutivas entre los organismos se pueden representar mediante árboles filogenéticos. Este tipo de árbol supone un ancestro común a todas las especies; es decir una expresión raíz para la cadena del 16S rRNA que se modifica al avanzar hacia las hojas. Éstas resultan ser las distintas secuencias que se hallan en la muestra metagenómica y, dentro de la estructura jerárquica del árbol, los agrupamientos pueden realizarse utilizando distintos niveles de similaridad o disimilaridad que se asocian con cada categoría taxonómica. Sin embargo es necesario aclarar que el árbol filogenético de secuencias del gen 16S rRNA no necesariamente reflejará con exactitud el árbol filogenético de los taxones que lo conforman pues pueden presentarse diferencias en las cadenas genéticas dentro de una misma especie. Hillis et al. [9]. De todas maneras, los grupos que se constituyan en cada nivel de disimilaridad pueden servir para evaluar de manera aproximada la diversidad en cada categoría taxonómica y con esta idea se los denomina Unidades Taxonómicas Operacionales (OTUs). Ver Schloss y Handelsman [10]

El primer paso para realizar el agrupamiento es alinear las distintas secuencias para establecer las zonas de mayor parecido y definir una “distancia genética” que, una vez alineadas, permita representar la diferencia entre dos cadenas de bases químicas. En la Figura 2.3.1 se muestra un ejemplo de alineamiento.

Figura 2.3.1

Alineamientos



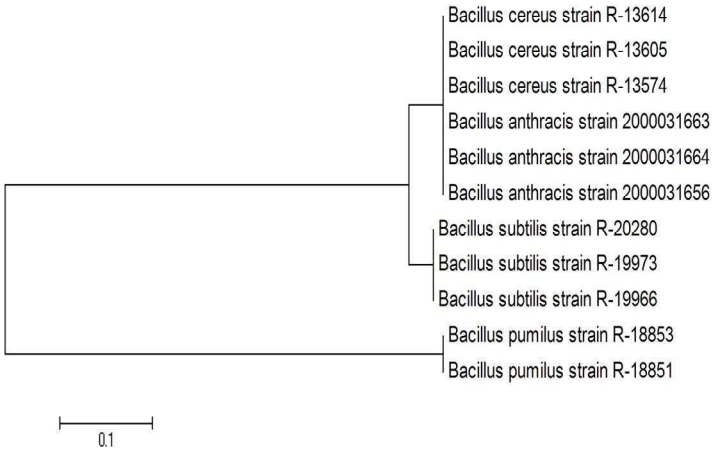
En su forma más simple la disimilaridad entre dos secuencias es igual al número de posiciones enfrentadas por el alineamiento que presentan distinto contenido. Por lo tanto una idea es usar la distancia de Hamming que se calcula como la proporción de celdas en las que la diferencia entre residuos ocurre [9]. Sin embargo hay algunas cuestiones de naturaleza biológica que sugieren cambios en la fórmula de las distancias a través de distintos modelos expuestos por Hillis et al. en [9]. En base a ellos se calcula la matriz de distancias utilizando todas las secuencias disponibles. Cada celda de la matriz es la distancia que hay entre la secuencia de la fila y la de la columna. Para los modelos usados habitualmente se han desarrollado programas aplicables on-line que la calculan, utilizando la distancia establecida en ellos. Una vez más conviene aclarar que dichas distancias solo pueden medirse una vez que las secuencias están alineadas, cuando ya se han establecido en forma general las zonas de mayor parecido.

A continuación se debe fijar el criterio con el que las secuencias se considerarán similares. El enfoque filogenético que requiere las distancias para la construcción del árbol hace usual el cálculo de las mismas para luego establecer las disimilaridades expresadas en porcentaje. En este punto es importante aclarar que también es habitual utilizar el porcentaje de disimilaridad complementario de la similaridad, lo que resalta la diferencia y no el parecido. En cualquier caso hay que

tener en cuenta que la taxonomía reconoce como jerarquías de orden creciente a especie, género, familia, orden, clase, phylum y dominio. Los distintos porcentajes de disimilaridad indican la máxima diferencia que se acepta entre cadenas de 16S rRNA correspondientes a los individuos de un grupo. Si bien en términos biológicos no puede aplicarse estrictamente el criterio computacional enunciado, es de práctica considerar que una disimilaridad de hasta el 3% corresponde a individuos de la misma especie mientras que para una disimilaridad que no exceda el 5% se considera igual género o que, para otra del 20%, hay igual clase o phylum. Las OTUs así obtenidas se citan subindicando el porcentaje referido: $OTU_{3\%}$ u $OTU_{20\%}$ por ejemplo. El árbol resultante es entonces un dendrograma que se corta al nivel del taxón requerido por el estudio. La Figura 2.3.2 ilustra esta situación.

Figura 2.3.2

Árbol Filogenético



La construcción de las OTUs puede realizarse de distintas formas. El criterio del "vecino más cercano" asigna una secuencia en una OTU cuando su disimilaridad con una cualquiera del grupo no supera el porcentual elegido. En cambio el método del "vecino más lejano" es más restrictivo pues asigna una secuencia en una OTU cuando su disimilaridad con cada una de las secuencias que la integran no supera el porcentaje definido. El procedimiento del "vecino promedio" utiliza la disimilaridad promedio entre las secuencias de una OTU y las que están fuera de él para asignar una secuencia cuando su disimilaridad con las del grupo resulta menor que dicho promedio. Cualquiera de estas variantes se construye sobre la teoría expuesta por Everitt en [11] y se lleva a cabo a través de programas específicos disponibles en la web.

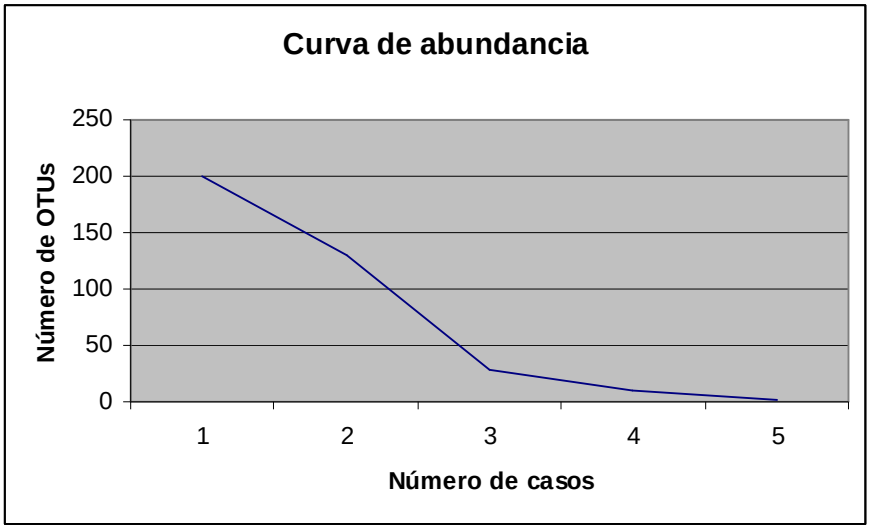
2.4 BIODIVERSIDAD

Para medir la diversidad de una comunidad biológica se tienen en cuenta diferentes conceptos: riqueza, abundancia, uniformidad y dominancia. Por riqueza se entiende aquí cantidad de taxones distintos presentes en el medio. Por ejemplo cantidad de especies diferentes. Abundancia se refiere a la cantidad de individuos correspondientes a cada taxón. Así, si todos los taxones presentes en la comunidad tuvieran igual cantidad de individuos habría uniformidad mientras que si algunos tuvieran mayor cantidad de individuos que otros, habría dominancia de esos taxones. En cualquier caso las mediciones resultan estimaciones estadísticas realizadas por medio de algún modelo matemático que trata de atenuar la incertidumbre producida por el gran volumen de datos y la incerteza sobre la cantidad de especies, o más generalmente taxones, presentes. En adelante, por constituir la circunstancia mas común, se hablará de especies sobreentendiendo que iguales conceptos pueden corresponder a distintos niveles taxonómicos

Los diferentes índices de riqueza elaborados a partir de estos modelos estiman el número de especies distintas, presentes en el medio, mientras que los modelos de abundancia establecen la distribución del número de veces que se presentan las especies en la comunidad. En la Figura 2.4.1 la curva de abundancia corresponde a la distribución de una muestra integrada por cadenas de 16S rRNA. Se encontraron 200 $OTUs_{3\%}$, con un solo individuo, 130 $OTUs_{3\%}$ con dos individuos, 28 $OTUs_{3\%}$ con tres representantes, 10 $OTUs_{3\%}$ con cuatro y 2 $OTUs_{3\%}$ con cinco individuos. El número total de cadenas analizadas es $200x1 + 130x2 + 28x3 + 10x4 + 2x5 = 594$ Sin embargo la cantidad de especies u $OTUs_{3\%}$ distintas es $200 + 130 + 28 + 10 + 2 = 370$ que resulta la riqueza de la muestra.

Figura 2.4.1

Representación Gráfica de la Abundancia



A efecto del modelado la abundancia puede presentarse también ordenando las frecuencias de aparición de cada OTU de mayor a menor $X_1, X_2, ..., X_S$. En la medida que el tamaño de la muestra crezca tales frecuencias tenderán a adoptar el valor de probabilidad de aparición de una OTU en una muestra de la población.

La distribución de las especies en la comunidad podría tener diversas formas según resulten los valores p_i correspondientes a la proporción sobre el total de individuos que cada una de las S especies presentes. En un extremo podría suponerse que en la comunidad hay la misma cantidad de organismos de cada especie. En el otro, la suposición sería que todos los organismos pertenecen a una misma y única especie. Se diría entonces que en el primer caso hay uniformidad de especies mientras que en el segundo hay dominancia de la única especie. Aún sin considerar estas situaciones extremas, el grado de la relación uniformidad-dominancia quedará establecido por algunas especies que aparezcan comúnmente y por otras que resulten raras. Al tomar una muestra aleatoria de la comunidad es posible entonces que ciertas especies, que debieran ser contabilizadas en su riqueza, no aparezcan por ser raras o al menos no las más comunes. ¿Cuál debe ser el tamaño de la muestra para asegurar la presencia en ella de todas o casi todas las especies? La respuesta no solo depende de la relación uniformidad-dominancia desconocida "a priori" sino también de las cantidades de especies posibles y de organismos en la comunidad que tampoco se conocen. Además

todas las cantidades y relaciones varían temporalmente lo que aporta mayor incertidumbre.

A partir de estas consideraciones es posible plantear la medición de la biodiversidad con diferentes enfoques. Se pueden utilizar índices de riqueza de especies, índices de uniformidad y dominancia y, más generalmente, modelos de abundancia de especies. En el primer caso se intenta medir solo la cantidad de especies presentes en el medio, mientras que en el segundo se trata de evaluar el grado de desorden distributivo que posee la comunidad. La tercera variante requiere un ajuste estadístico de la curva de abundancia. Cualquiera de las tres alternativas ofrece una gran variedad de medidas que presentan valores muy diferentes y plantean por ello problemas acerca de su interpretación dentro de un medio biológico. Esto dificulta, a su vez, las comparaciones de biodiversidad entre comunidades.

Dos medidas comúnmente usadas son: el índice S de riqueza de especies y la medida E de uniformidad de especies.

El índice de riqueza S establece simplemente el número de especies que hay en el medio. Como analiza O'Hara en [12], su cálculo puede hacerse por estimación no paramétrica, por estimación paramétrica de máxima verosimilitud a partir de la curva de abundancia, o utilizando curvas de rarefacción. Estas últimas se basan en procedimientos de remuestreo usuales en explotación de datos.

El índice de uniformidad E utiliza el concepto de entropía definido en teoría de la información. Ver Shannon [13] La entropía se calcula de acuerdo a

$$H = -\sum p_i \ln p_i \qquad (2.4.1)$$

dónde p_i es la probabilidad de ocurrencia de la i-ésima especie. Si hay uniformidad de especies, es decir si todas las especies tienen la misma probabilidad de ser observadas, la entropía es máxima. De tal forma para una cantidad $S = n$ de especies resultaría

$$H = -\sum_{i=1}^{n} \frac{1}{n} \ln \frac{1}{n} = -n \frac{1}{n} \ln \frac{1}{n} = -\ln \frac{1}{n} \qquad (2.4.2)$$

Para definir un índice normalizado Hill et al. toman en [14]:

$$E = \frac{H}{\ln S} \qquad (2.4.3)$$

y en este caso, si hay uniformidad, queda

$$E = \frac{-\ln\frac{1}{n}}{\ln n} = \frac{-\ln\frac{1}{n}}{\ln\left(\frac{1}{n}\right)^{-1}} = \frac{-\ln\frac{1}{n}}{-\ln\frac{1}{n}} = 1 \qquad (2.4.4)$$

Si la distribución de especies en la población se va deslizando desde la uniformidad hacia la dominancia el índice E se moverá consecuentemente hacia 0. En efecto; en la situación teórica extrema en que exista una sola especie resulta

$H = -p_{especie} \ln p_{especie} = -1\ln 1 = 0$ Y entonces también $E = 0$

Existen otros índices y medidas para evaluar la forma de la distribución. En [15] Magurran cita el índice de Simpson como una de las medidas de mayor significación y más robustas disponibles. Sin embargo en el presente trabajo se utilizará la entropía calculada como H según (2.4.1), pues es la más difundida en el ámbito computacional y porque se intenta analizar con ella el desempeño de los algoritmos propuestos aclarando su significación en ese contexto.

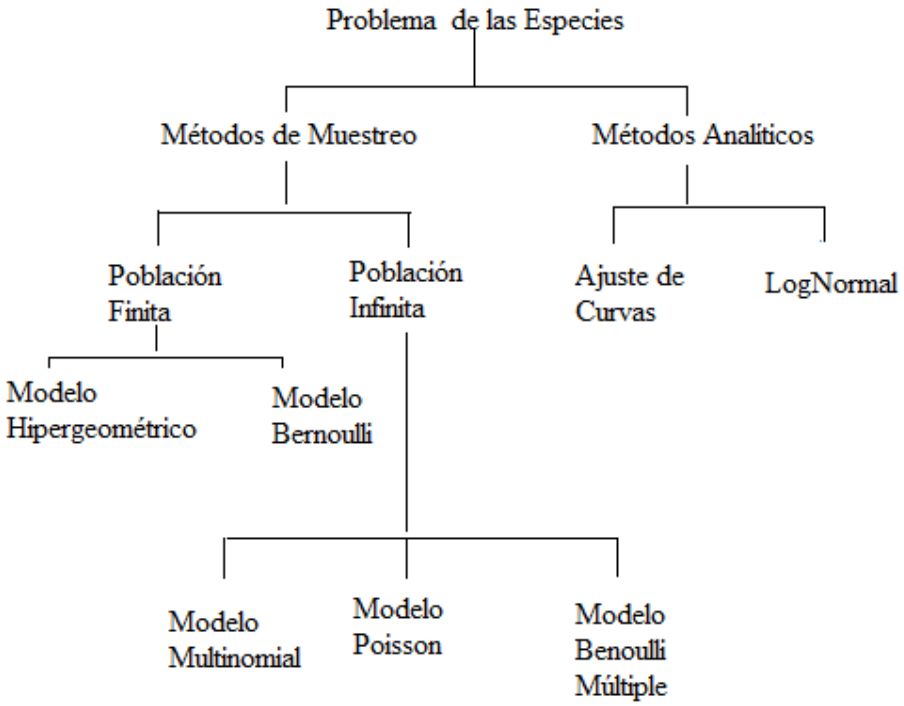
3- ESTADO DEL ARTE

3.1 EL MODELO MULTINOMIAL Y LA ESTIMACIÓN DE LA RIQUEZA

La diversidad biológica de una comunidad es función conjunta de dos conceptos: la riqueza, entendida como la cantidad de especies (o de taxones en general) y la distribución de las especies (o taxones) en su sentido estadístico. Desde 1932 en adelante se han realizado esfuerzos para modelar la relación entre el número de especies y la cantidad de individuos de cada una de esas especies en una dada comunidad, desarrollándose así los llamados modelos de abundancia. Ver Magurran [15]. Los modelos matemáticos utilizados conducen a distintas formas de evaluación de la riqueza y distribución. Bunge y Fitzpatrick presentan en [16] los diferentes criterios de trabajo que pueden aplicarse atendiendo a las características de cada tipo de comunidad. La Figura 3.1.1 brinda un esquema de esas posibilidades.

Figura 3.1.1

Enfoques para Análisis de Biodiversidad



En los estudios metagenómicos, el análisis de la biodiversidad procura establecer la riqueza como uno de los principales parámetros poblacionales. La riqueza es entonces una cantidad que debe inferirse a partir de muestras, lo que suele presentar problemas pues es difícil que se posea una idea previa del número de individuos que forman la población y muy poco frecuente que se tenga una clara noción de cómo esos individuos se distribuyen entre las distintas especies.

La distribución en el medio puede presentar algunas especies dominantes, otras que no lo son tanto y una mayoría que en términos estadísticos resultan raras. Ver Magurran [15]. De modo que al elegir al azar un individuo de la comunidad, la probabilidad p_i de que sea de la especie i puede ser alta, media o muy baja según el grado de dominancia o rareza que tenga la especie. Al tomar una muestra de la comunidad se selecciona una cantidad x_i de individuos de esa especie y, suponiendo independencia en la elección de cada individuo, se tiene un vector $(x_1, x_2, ..., x_S)$ con las cantidades presentes de cada especie. La probabilidad de elegir esa muestra se distribuye en forma multinomial según:

$$p(x_1, x_2, ..., x_S) = \binom{n}{x_1, x_2, ..., x_S} p_1^{x_1} p_2^{x_2} ... p_S^{x_S} \tag{3.1.1}$$

donde S es la cantidad de especies

$$n = x_1 + x_2 + ... + x_S \tag{3.1.2}$$

y

$$\binom{n}{x_1, x_2, ..., x_S} = \frac{n!}{x_1! x_2! ... x_S!} \tag{3.1.3}$$

Ver Chao y Shen [17]

El problema se plantea cuando para la muestra escogida algunos valores de x_i son cero, pues esto significa en la práctica que no se va a tener registro de la presencia de esas especies en la comunidad. Es decir; como la cuenta de la cantidad de especies se hace a partir de las que están presentes en la muestra, resulta que se estiman menos especies de las que realmente hay. En particular es claro que una especie rara tiene menos probabilidad de figurar en la muestra que una dominante. La respuesta teórica de la estadística frente a esta situación es aumentar el tamaño de la muestra para lograr captar también casos de especies raras. Pero esto puede no ser posible ni razonable. En efecto; supóngase que se tiene una especie i cuya proporción en la

población es uno de cada diez mil individuos; entonces $p_i = \frac{1}{10000}$ es la probabilidad de que al elegir un individuo de la comunidad este pertenezca a la especie i. El biólogo desconoce la presencia de esta especie rara en la comunidad y desearía tomar una muestra cuyo tamaño le asegure encontrarla. Para calcular el tamaño de la muestra hay que ponerse en la situación binomial de elegir con independencia un individuo correspondiendo éxito si es de la especie en cuestión y fracaso si no. Un intervalo de confianza $(1-\alpha)\%$ para estimar tal probabilidad a partir de la proporción muestral es:

$$(\hat{p}_i - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}, \hat{p}_i + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}) \tag{3.1.4}$$

Por lo tanto al fijar un error ε para la estimación se tiene:

$$n = \frac{z_{\frac{\alpha}{2}}\hat{p}_i(1-\hat{p}_i)}{\varepsilon^2} \tag{3.1.5}$$

Una estimación suficientemente próxima de p_i debería ser $\hat{p}_i \approx \frac{1}{10000} = 0.0001$ y admitiendo un error $\varepsilon = 0.00005$ para una confianza $(1-\alpha)\% = 99\%$ se tiene $n = \frac{(2.58)^2 \times 0.0001 \times 0.9999}{(0.00005)^2} \approx 266229$ Este es el tamaño de muestra que con un 99% de confianza captaría una proporción del orden del diezmilésimo con un error en más o en menos de cinco cienmilésimos. Lo que equivale a decir que, con esta cantidad de individuos en la muestra, casi con certeza se hallaría la especie i.

Más allá del aspecto tecnológico, si se tienen en cuenta costos, hábitos de trabajo y dimensiones de muestras ya almacenadas en las distintas bases, un tamaño como el indicado resulta poco menos que imposible. Si se adopta el punto de vista del análisis de "marcadores", para obtener unos pocos miles de secuencias del gen 16S rRNA, cada una de las cuales representa a un individuo, hace falta procesar una gran cantidad de datos. Por lo tanto los tiempos, las características de la recolección de las muestras físicas, las velocidades de procesamiento de las computadoras y los costos involucrados en obtener muestras del tamaño propuesto, impiden en los hechos contar con ellas.

La conclusión es que el modelo estadístico planteado es explicativo conceptualmente pero la estimación de sus principales parámetros resulta una difícil tarea. Como señalan Hughes, J et al en [18], la alta diversidad y, en relación con ella,

el tamaño pequeño de la muestra, se combinan para abonar la idea de que la riqueza microbiana no puede ser estimada adecuadamente, al menos si se utilizan las técnicas estadísticas usuales.

Para lograr estimaciones más adecuadas se han sostenido distintas líneas de trabajo relacionadas con los aspectos propiamente estadísticos y también con los procesos que, previos a la estimación, se aplican a las secuencias de ADN.

3.2 ESTIMACIONES NO PARAMÉTRICAS

Desde el punto de vista del mejoramiento de la técnica estadística Ann Chao desarrolló en [19] un nuevo estimador de la riqueza de especies.

$$S_{Chao1} = D + \frac{f_1^{\,2}}{2(f_2+1)} - \frac{f_1 f_2}{2(f_2+1)^2} \tag{3.2.1}$$

dónde f_r es la cantidad de especies que aparecen r veces en la muestra de tamaño

$$n = \sum_{r=1}^{r\max} r f_r$$

(3.2.2)

$$D = \sum_{r=1}^{r\max} f_r \tag{3.2.3}$$

es la cantidad total de especies observadas y claramente si S es la cantidad desconocida de especies en el medio resulta

$$D = S - f_0 \tag{3.2.4}$$

con f_0 número de especies que no aparecen en la muestra, también desconocido.

La varianza de esta estimación es:

$$V_{Chao1} = f_2\left(\frac{1}{4}\left(\frac{f_1}{f_2}\right)^4 + \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{2}\left(\frac{f_1}{f_2}\right)^2\right) \tag{3.2.5}$$

O'Hara destaca en [12] que la cantidad así calculada funciona bien como estimador siempre que sea considerada como la estimación de una cota inferior.

Con el fin de mejorar la estimación Chao y Lee [20] construyeron un nuevo estimador basado en la idea de cobertura de la muestra. En el artículo, dicen los autores: "*Si estimamos el número de clases sin estimar la variación entre las probabilidades de las clases, normalmente solo podremos estimar una cota inferior, siendo la cota inferior producida en el caso equiprobable*", es decir, el caso en que

todas las especies tienen la misma probabilidad de ser elegidas y el medio es uniforme. Al respecto debe observarse que en tal situación la entropía se calcula:

$$H = -\sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{n}) = -n \frac{1}{n} [\log 1 - \log n] = \log n \qquad (3.2.6)$$

siendo n el número de especies distintas. Por lo tanto:

$$e^{H} = n \qquad (3.2.7)$$

da el número de especies presentes cuando su distribución es uniforme. Si para un dado valor de H la distribución no es uniforme el número real de especies tendrá que ser mayor que esta estimación, que funcionaría como cantidad mínima de especies que debieran esperarse.

La cobertura se obtiene como suma de las probabilidades de las especies. Debe crecer cuando mas especies están en la muestra, pero lo hace en relación con la proporción que cada especie va teniendo. La idea es que también tiene en cuenta las variaciones de la probabilidad además de la porción de la distribución cubierta. Su estimación se realiza por medio de:

$$\hat{C} = 1 - \frac{f_1}{n} \qquad (3.2.8)$$

donde f_1 es el número de especies representadas en la muestra por un solo individuo. La fracción

$$T = \frac{f_1}{n} \qquad (3.2.9)$$

es, según lo informa Good en [21], una cantidad sugerida por Alan Turing para estimar la probabilidad de descubrir una especie nueva al agregar un nuevo individuo a la muestra. Chao y Lee utilizan en [20], además, una estimación del coeficiente de variación γ de las probabilidades de las clases, lo que conduce a la fórmula del estimador de riqueza ACE (Abundance Coverage Estimator):

$$S_{ACE} \approx \frac{E(D)}{E(C)} + \frac{E(f_1)}{E(C)} \gamma^2 \qquad (3.2.10)$$

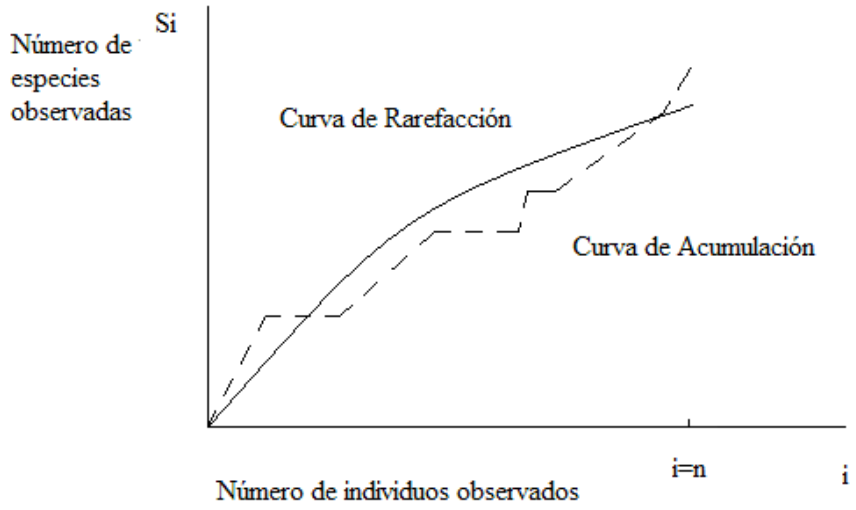
3.3 CURVAS DE RAREFACCIÓN

Un enfoque alternativo a la estimación no paramétrica de la riqueza, es el que aportan las curvas de rarefacción. El método permite estimar la riqueza de un medio aunque generalmente es aplicado para comparar riqueza entre dos o más comunidades,

pues alivia los problemas derivados del tamaño insuficiente y habitualmente desigual de las muestras [22]. Hay dos variantes: rarefacción por individuos, que se basa en la recolección de una sola muestra del medio, y rarefacción por muestras para la cual es necesario tomar varias de ellas distribuidas en forma espacial o temporal [23]. La idea básica al tratar de estimar la riqueza es que a partir de la toma de muestras de mayor tamaño será posible capturar un número creciente de especies distintas. Dada una comunidad que tenga una cantidad desconocida N de individuos y un número S de especies distintas también desconocido, se pueden tomar muestras de tamaño n y determinar S_n que es el número de especies distintas halladas en una muestra. El valor esperado teórico de $E(S_n)$ se aproxima por el promedio de los S_n y se utiliza para medir la riqueza S del medio. En el caso de rarefacción por individuos se comienza construyendo una curva de acumulación del número de especies distintas según se ve en la gráfica punteada de la Figura 3.3.1

Figura 3.3.1

Acumulación y Rarefacción



La curva punteada muestra la acumulación del número de especies distintas conforme se van examinando cada uno de los n individuos que forman la muestra. El procedimiento de rarefacción aplica una técnica de remuestreo planteada por Efron en [24]. Consiste en repetir muchas veces este análisis de individuos,

tomando cada vez un orden distinto y aleatorio de los n casos y estableciendo el número acumulado promedio para cada cantidad i de casos examinados. La curva resultante tiene un aspecto suave como se observa en la línea llena de la Figura 3.3.1. Es decir; la curva de rarefacción representa el promedio de todas las curvas de acumulación construidas.

Si se llama S_{iobs} al valor de ordenadas de la curva de rarefacción empíricamente construida en cada valor i, pueden establecerse intervalos de confianza para S_i. Obsérvese que este valor esperado resultaría el promedio S_i de las cantidades de especies, tomado sobre todos los órdenes posibles en que los individuos pudieran ser considerados. Como este número es $i!$ y como debiera considerarse un nuevo reordenamiento completo cada vez que se incrementa en una unidad la cantidad de individuos ($i = 1,...,n$), se comprende la necesidad de considerar, para cada valor de i, solo algunos reordenamientos de la muestra. Por esta razón S_i se estima por intervalos de confianza. En efecto, para el 95% de confianza el intervalo respectivo es: $(S_{iobs} - 1.96\sigma_{S_{iobs}} \leq S_i \leq S_{iobs} + 1.96\sigma_{S_{iobs}})$

Cuando $i = 1$ necesariamente hay una sola especie y por lo tanto $\sigma_{S_{1obs}} = 0$ mientras que si $i = n$ cualquier remuestreo producirá el mismo valor que el promedio S_n y por lo tanto también en este caso $\sigma_{S_{nobs}} = 0$. De tal modo se obtiene, para un tamaño n de la muestra, una estimación S_n.

Si se tomaran en la comunidad a su vez varias muestras de tamaño n, los distintos valores de S_n tendrían una distribución en el muestreo con valor esperado $E(S_n)$. En términos teóricos $E(S_n)$ es un estimador asíntótico de S, la riqueza de la comunidad. Podría entonces recurrirse al incremento de la cantidad de elementos en la muestra a fin de lograr una estimación adecuada de la riqueza poblacional. Por desgracia esto no es operativamente posible, pues la cantidad de individuos necesarios para que la muestra sea representativa es desconocida y seguramente muy grande para las posibilidades de la tecnología de secuenciación y de proceso de los datos [22]. Sin embargo, si se utilizan curvas de rarefacción para evaluar la riqueza, es posible establecer una asíntota horizontal cuyo valor resulte al menos una aproximación a la riqueza del medio. Para ello debe contarse con una expresión analítica de la curva que permita extrapolar su comportamiento más allá de los tamaños de muestras posibles y determinados. En la práctica el método de rarefacción por individuos implica entonces

tomar una muestra lo mas grande posible y realizar el promedio de algunas curvas de abundancia incorporando de a uno los individuos al análisis. A partir de aquí se halla una expresión analítica para la curva de rarefacción por medio de algún procedimiento de ajuste. En particular puede realizarse un ajuste de mínimos cuadrados utilizando una hipérbola rectangular según lo plantea Tellinghuisen en [25]. Tal hipérbola dada por la expresión general

$$\hat{S}_n = \frac{Vn}{K+n} \tag{3.3.1}$$

es llamada, en el contexto biológico, curva de Michaelis-Menten y la exactitud con la que pueden estimarse sus parámetros V y K es analizada por Currie en [26]. Como un camino alternativo O'Hara cita en [12] el uso de una curva exponencial. Por cualquier vía que se llegue a ella, una vez hallada la expresión matemática, se extrapola el comportamiento asintótico y se halla así un valor de $\hat{S}_n$ que aproxime suficientemente a S, la riqueza de la comunidad.

Varias curvas de rarefacción calculadas para muestras de distinto tamaño permiten comparaciones de riqueza según el punto a partir del cual se considera que se alcanzó el valor de la asíntota horizontal. Pero a la vez, la rapidez con que esa asíntota sea alcanzada en cada caso, puede permitir una apreciación de la diversidad entendida como diferencia de abundancias. En efecto; es dable esperar que en un medio uniforme, dónde todas las especies tengan parecida probabilidad de ser halladas, se requieran menos individuos, para que la curva alcance su asíntota, que en aquel medio donde exista mayor diferencia en la distribución y haya especies con muy baja probabilidad de aparecer en una muestra.

3.4 PROCESOS PREVIOS A LA ESTIMACIÓN

En relación con los procesos que se efectúan sobre las secuencias una vez obtenidas del secuenciador y que resultan previos a la estimación de biodiversidad se han observado distintos efectos que influyen luego sobre la determinación de la riqueza. El modelo utilizado para agrupar las secuencias en taxones requiere alinearlas de acuerdo a un patrón, filtrarlas para eliminar los "gaps" que en ellas se presentan y finalmente calcular las distancias de cada una de ellas a todas las demás que integran la muestra, de acuerdo a una definición de "distancia genética" que permita representar la diferencia entre cadenas de bases químicas. Al respecto se trabaja sobre la idea de

distancia de Hamming que calcula la cantidad de celdas enfrentadas con distinto contenido. Así se cumple con los requerimientos matemáticos para establecer una distancia entre dos secuencias x e y:

i) $d(x, y) \geq 0$

ii) $d(x, y) = 0 \Leftrightarrow x = y$

iii) $d(x, y) = d(y, x)$

iv) Si z es una tercera secuencia

$$d(x, z) \leq d(x, y) + d(y, z)$$

Sobre esta base se consideran, no obstante, cuestiones de tipo biológico que modifican y complementan el cálculo. En primer término el marco teórico en el que se establece la noción de distancia es el de un modelo probabilístico de la evolución, es decir de la mutación y selección, eventos que fueron cambiando las secuencias de ADN, lo que se esquematiza por medio de un árbol construido utilizando la distancia definida e inferido estadísticamente a partir de los datos. Ese modelo se realiza bajo el supuesto markoviano que considera cada posición en la cadena independiente de toda otra y dependiente solo del estado anterior de la misma. Dicho proceso se postula además estacionario en sentido amplio lo que significa que la probabilidad de sustitución de una base química de la cadena por otra depende sólo del tiempo transcurrido entre que se presenta una y su sustituta. Estas son hipótesis simplificadoras que facilitan el modelado.

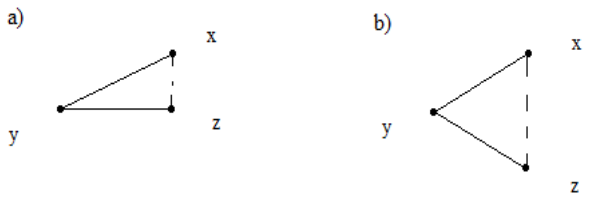
En segundo lugar ocurre que, en una dada posición de las secuencias alineadas, pueden presentarse las mismas bases sin que esto implique que en el pasado evolutivo haya ocurrido lo propio. Se trataría entonces de una superposición de sustituciones en una de ellas o en ambas, que llevó al mismo nucleótido. Si este fuera el caso para varias posiciones de las cadenas, la igualdad actual de residuos no debiera garantizar la igualdad en cuanto a la OTU considerada, pues las secuencias estarían realmente más lejanas de lo que parece en la evolución. Además la superposición de sustituciones no crece uniformemente con el número de eventos evolutivos sino que lo hace más rápidamente al principio y más lentamente luego. No obstante todo ello, el principio de parsimonia, aplicado en la estructuración del árbol, implica la consideración del mínimo de cambios necesarios en cada sitio para explicar la vinculación entre las secuencias. Por lo tanto surge la necesidad de corregir el cálculo a efecto de representar la verdadera distancia evolutiva [27].

Como tercer elemento hay que considerar que durante la evolución del gen pueden hallarse “inserciones” de nucleótidos o “deleciones” de los mismos que obligan a intercalar sitios vacíos o “gaps” para comparar alineamientos. En verdad, la práctica común es omitir los tramos con gaps en el análisis, eliminando directamente de ambas secuencias las posiciones que resulten vacías en al menos una de ellas o, mas drásticamente, eliminando esos sitios para todas las secuencias que se estudian.

El cuarto aspecto para analizar es la forma conceptual del árbol filogenético pues ella determina cierta propiedad deseable de la distancia entre organismos. En efecto; la forma del árbol podría buscarse a través del procedimiento UPGMA [3] y en tal caso se asumirá que el tiempo en que se produce la variación de una secuencia para formar otra se mide con un “reloj molecular” que asigna un lapso constante a este proceso. Esto es equivalente a considerar que la distancia, a través de los arcos, desde un nodo cualquiera a todas las hojas que dependen de él, es la misma. Por supuesto no necesariamente debe aceptarse esta hipótesis, por lo cual el árbol puede construirse también por otros procedimientos que no la tengan en cuenta. Pero, si se la considera, el árbol filogenético resultante poseerá la propiedad ultramétrica. Esta propiedad es matemáticamente más fuerte que la desigualdad triangular citada en iv) pues para tres secuencias distintas x, y, z se requiere que $d(x,z) \leq Max\{d(x,y), d(y,z)\}$ La Figura 3.4.1 ilustra la diferencia entre ambas propiedades. El triángulo a) cumple la propiedad triangular en cualquier orden en que se tomen los puntos. Por ejemplo $d(y,x) \leq d(y,z) + d(z,x)$. Sin embargo no se verifica la propiedad ultramétrica ya que $d(y,x) \succ Max\{d(y,z), d(z,x)\}$. En cambio en el triángulo b) se cumple la desigualdad triangular pero además, para cualquier orden de los puntos que se asuma, también se verifica la desigualdad del máximo. Por ejemplo $d(y,x) \leq Max\{d(y,z), d(z,x)\}$.

Figura 3.4.1

Propiedades de las Métricas



Todas las consideraciones realizadas están presentes en los supuestos del modelo de evolución y condicionan la expresión que adopta la distancia genética modificando la fórmula original de Hamming. El marco general del modelo evolutivo está dado por una matriz denominada de "divergencia" en cuyas celdas figuran las proporciones de pares de las secuencias alineadas x e y que cambian de un nucleótido a otro:

$$F_{xy} = \begin{bmatrix} N_{AA}/N & N_{AC}/N & N_{AG}/N & N_{AT}/N \\ N_{CA}/N & N_{CC}/N & N_{CG}/N & N_{AT}/N \\ N_{GA}/N & N_{GC}/N & N_{GG}/N & N_{GT}/N \\ N_{TA}/N & N_{TC}/N & N_{TG}/N & N_{TT}/N \end{bmatrix} \tag{3.4.1}$$

dónde N_{IJ} es la cantidad de veces que se cuenta el par IJ de bases en las secuencias alineadas y el total de pares está dado por $N = \sum N_{IJ}$. A efecto de simplificar la notación se utiliza la representación de las distintas fracciones por letras según:

$$F_{xy} = \begin{bmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{bmatrix} \tag{3.4.2}$$

A partir de aquí puede definirse la disimilaridad por medio de

$$D = 1 - (a + f + k + p) \tag{3.4.3}$$

fórmula que resta del total la suma de los elementos de la diagonal de la matriz que indican las proporciones de ausencia de cambios. Sin embargo llegado este punto pueden plantearse distintas alternativas para evaluar la distancia entre secuencias que atiendan a las consideraciones realizadas. En la forma adoptada por Jukes-Cantor se supone que los elementos de la diagonal cuando el tiempo de evolución crece van tendiendo a igualarse y lo propio ocurre entre los elementos que no están en la diagonal matricial. Se define entonces la distancia entre dos secuencias x e y como:

$$d_{xy} = -\frac{3}{4}\ln(1 - \frac{4}{3}D) \tag{3.4.5}$$

Se observa aquí que la disimilaridad D esperada es menor que $\frac{3}{4} = 0.75$ pues con este valor ya no podría calcularse el logaritmo. Con esta salvedad la función definida cumple con las condiciones exigidas matemáticamente para una distancia. Se han planteado distintas variantes que toman en cuenta otras suposiciones más elaboradas no obstante lo cual es de práctica el uso de esta distancia por su expresión simple y su fácil cálculo [27]. La relación entre la disimilaridad y la distancia calculada queda establecida por la fórmula dada mas arriba o equivalentemente por

$$D = \frac{3}{4}(1 - e^{-\frac{4}{3}d_{xy}}) \qquad (3.4.6)$$

Con el modelo de Jukes-Cantor y sus distintas modificaciones se han desarrollado programas que calculan la matriz de distancias para luego agrupar las secuencias en diferentes OTUs. Ver Schloss y Handelsman [28].

En [29] Schloss analiza como la calidad del alineamiento, el método de cálculo de la distancia genética, el filtro aplicado a las secuencias y la región de análisis elegida en el gen 16S rRNA impactan sobre las mediciones. En primer lugar algunos patrones utilizados tienen un pobre desempeño en el alineamiento. Las diferencias entre las bases de datos patrones más comúnmente utilizadas, tales como Silva, Greengenes y RDP, en cuanto a la cantidad de secuencias en cada una, a la atención puesta en la creación del alineamiento y a la calidad del curado, influyen en los cálculos de las distancias entre pares de secuencias. Un segundo aspecto surge de los distintos modelos que se emplean para calcular la distancia genética pues, en realidad, las secuencias utilizadas son sólo una parte de la cadena total del gen. En tercer término, el filtrado que se realiza para tratar los “gaps” parte de la idea de considerarlos como datos desaparecidos en la operación electroquímica de secuenciación y no como reales mutaciones genéticas, lo cual podría no ser siempre cierto. Finalmente el cuarto punto es el que resulta de la práctica de remover también las regiones consideradas variables del gen porque son las que menor conservación presentan a través del proceso evolutivo. Todos estos efectos se asocian para determinar la necesidad de reducir los valores de disimilaridad que identifican individuos de un mismo taxon [29].

3.5 ANÁLISIS DEL PROBLEMA

A fin de ejemplificar la disparidad y, en general, la subestimación que presentan las mediciones de riqueza como consecuencia de los aspectos que se han venido señalando en las secciones anteriores, se expone el caso analizado por Roesch et al en [30]. Se trata de una colección de 2702 secuencias del gen 16S rRNA obtenida de Ribosomal Database Project II de las cuales se sabe previamente que pertenecen a 2410 especies distintas incluidas en 685 géneros. Los resultados sobre las estimaciones de riqueza realizadas a partir de los datos por tres formas de estimación comúnmente utilizadas arrojaron los valores que se listan en la Tabla 3.5.1

Tabla 3.5.1

Disimilaridad %	Rarefacción	Chao1	ACE
0	1447	4358	5254
3	902	1700	1725
5	689	1122	1113
10	416	543	547

Si se considera por ejemplo que el taxón especie se corresponde aproximadamente con un nivel de disimilaridad del 3% se ve claramente la subestimación que de la cantidad real de especies (2410) producen los distintos modelos. Lo propio ocurre si se considera 10% de disimilaridad como indicativo de género.

Resultados como el citado obligan a buscar alternativas que permitan afinar la estimación y delimitar el significado de los valores hallados al relacionarlos con la forma de la distribución de especies y el esfuerzo de muestreo. Parece necesario entonces elaborar algunas herramientas adicionales que tengan en cuenta el enorme volumen inicial de los datos muestrales y su reducción posterior a cantidades que resultan insuficientes para la estimación. Tales herramientas podrán recurrir a procedimientos que usualmente acompañan a la minería de datos y se aplican para explorarlos, acompañados de mediciones y evaluaciones de error efectivas.

El primer paso en esa dirección, desde el punto de vista estadístico, puede darse al evaluar la probabilidad de que, en una muestra de determinado tamaño, se encuentre una especie cuya proporción en la población sea pequeña y por ende se la juzgue rara. Al considerar una muestra de tamaño $n = 6000$ se calcula la probabilidad de hallar una especie rara que está, por ejemplo, en proporción uno en diez mil en la población. La especie está (éxito) o no aparece en la muestra (fracaso) y por lo tanto la

probabilidad de hallarla al menos una vez en ella se obtiene de: $P(k \geq 1) = 1 - P(0) = 1 - \binom{6000}{0}\left(\frac{1}{10000}\right)^0\left(\frac{9999}{10000}\right)^{6000}$ Calculando:

$r = \left(\frac{9999}{1000}\right)^{6000}$, $\log r = 6000 \log \frac{9999}{10000} = -0.2606$ y $r = 10^{-0.2606} = 0.5488$ Finalmente entonces $P(k \geq 1) = 1 - 0.5488 = 0.4512$. En suma la probabilidad de detectar una especie rara con la proporción señalada en la población, al tomar una muestra de 6000 individuos es 0.4512. A continuación se muestra la Tabla 3.5.2 que revela la forma del crecimiento de la probabilidad de detección de acuerdo a la rareza de la especie conforme aumenta el tamaño de la muestra.

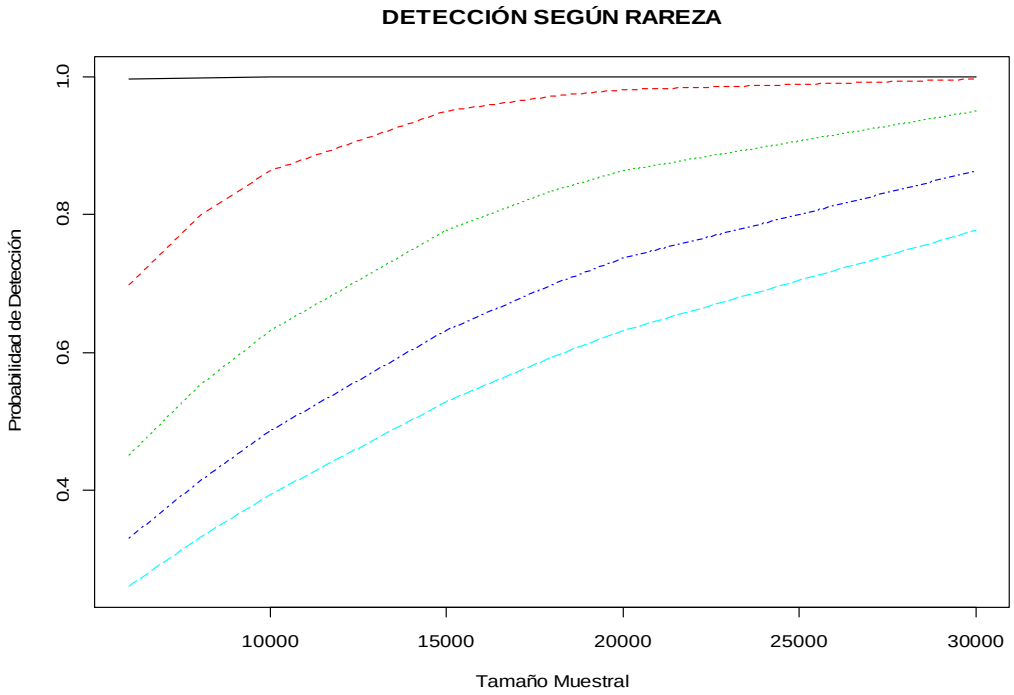
Tabla 3.5.2

Tamaño de la muestra	Probabilidad de detección para una proporción poblacional de $p_i = \frac{1}{1000}$	Probabilidad de detección para una proporción poblacional de $p_i = \frac{1}{5000}$	Probabilidad de detección para una proporción poblacional de $p_i = \frac{1}{10000}$	Probabilidad de detección para una proporción poblacional de $p_i = \frac{1}{15000}$	Probabilidad de detección para una proporción poblacional de $p_i = \frac{1}{20000}$
6000	0.9975	0.6988	0.4512	0.3297	0.2592
8000	0.9996	0.7981	0.5507	0.4113	0.3297
10000	1	0.8647	0.6321	0.4866	0.3937
15000	1	0.9502	0.7769	0.6321	0.5276
18000	1	0.9727	0.8347	0.6988	0.5934
20000	1	0.9817	0.8647	0.7364	0.6321
30000	1	0.9975	0.9502	0.8647	0.7769
266229	1	1	1	1	1

La información se exhibe en la Figura 3.5.3 que muestra las curvas de probabilidad de detección para cada proporción de rareza.

Figura 3.5.3

La Probabilidad de Detección



Las curvas desde la parte inferior de la Figura 3.5.3 hacia arriba corresponden al aumento de la proporción en la cual la especie se encuentra en la población.

3.6 CONCLUSIONES

La biodiversidad engloba los conceptos de riqueza y distribución de taxones que están relacionados. Para estimar la riqueza se requiere sortear dificultades de orden estadístico, a la vez que otros problemas ligados al modelado de los hechos biológicos y a su representación computacional. En el aspecto propiamente estadístico, se han construido distintos métodos tales como estimaciones no paramétricas y curvas de rarefacción, que arrojan disparidad en las mediciones y que subestiman los valores reales. Por otro lado, resulta claro que la probabilidad de detectar una mayor cantidad de especies, aún las técnicamente “raras”, crece con el aumento del tamaño muestral. Todos estos elementos constituyen la base del trabajo realizado que se propuso mejorar la estimación de la riqueza poblacional.

4. LOS DATOS Y EL PROCESO ESTÁNDAR

4.1 CONJUNTOS DE MUESTRAS

Se seleccionaron dos conjuntos de muestras a efecto de probar en ellas el desempeño de las estimaciones de riqueza realizadas por medio de las técnicas disponibles usualmente y de compararlas, luego, con los resultados obtenidos a partir de las ideas y mejoras propuestas.

El primer conjunto corresponde al suelo de La Sal del Rey, región lacustre hipersalina de baja profundidad, ubicada en el Estado de Texas, EEUU, cuyas coordenadas son (26° 31’ 55’’ N, 98° 03’ 50’’ O). Hollister et al [31] extrajeron material en 8 puntos distintos, a lo largo de una transecta de 140 metros, y obtuvieron ocho muestras integradas por secuencias de ADN correspondientes al gen 16S rRNA. Estas secuencias se encuentran almacenadas en NCBI Short Read Archive bajo el número de acceso SRX008158 [32], de donde fueron tomadas para desarrollar el trabajo. La denominación de cada muestra y la cantidad de secuencias que la integra se da en la Tabla 4.1.1 mientras que el número total de bases químicas almacenadas es de 925673.

Tabla 4.1.1

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
Tamaño	1641	8361	6926	6146	6226	8444	6103	5885

El segundo conjunto corresponde a suelo de la Selva Amazónica, en Brasil, con tres tipos de manejos. Sus coordenadas son (3° 26’ S, 60° 23’ O) y el número de acceso en NCBI Short Read Archive es ERX009564 [32]. Está constituido por seis muestras integradas por secuencias de ADN del gen 16S rRNA cuyo tamaño y nomenclatura se muestra en la Tabla 4.1.2. El total del conjunto es de 2400000 bases.

Tabla 4.1.2

Muestra	Err19	Err20	Err21	Err22	Err23	Err24
Tamaño	5011	5582	7637	3299	10371	5840

4.2 PROCESOS INICIALES

Los procesos efectuados sobre las secuencias contenidas en ambos conjuntos de muestras fueron los necesarios para establecer mediciones de biodiversidad. Consistieron en el alineamiento de secuencias, el filtrado de las mismas de acuerdo a su calidad, la determinación de la matriz de distancias y finalmente el agrupamiento en unidades taxonómicas operacionales (OTUs). Esto se realizó utilizando el software libre MOTHUR desarrollado por el equipo dirigido por Patrick Schloss [33]. Se detallan a continuación las características principales de cada proceso aplicado a las muestras.

Las distintas secuencias son alineadas contra un alineamiento de referencia, que en el caso del gen 16S rRNA tiene en cuenta no solo el alineamiento óptimo, sino también la estructura tridimensional de la molécula. En forma general se busca el patrón más cercano a cada secuencia y se alinea con él la secuencia candidata. Los datos que utiliza MOTHUR son compatibles con la base de datos Greengenes, habitualmente usada para alinear secuencias del gen 16S rRNA. En este caso, para ambos conjuntos de muestras SRX008158 y ERX009564, el proceso de alineamiento se realizó en la forma más sencilla posible introduciendo sólo los parámetros relativos al conjunto de secuencias candidato y al patrón.

Al alinear es usual que queden “huecos” o “gaps” en distintas posiciones de las secuencias. En el listado de las mismas, esto corresponde a guiones o puntos colocados por ausencia del símbolo correspondiente a alguna base química. Tales “gaps” deben ser removidos antes de calcular la distancia entre secuencias, razón por la cual se las filtra. Para los dos conjuntos de muestras utilizados en este trabajo el filtrado se realizó en las condiciones “default”, que solo requieren en MOTHUR el nombre del archivo de entrada proveniente del proceso de alineado.

Para calcular las distancias entre secuencias se utilizó el programa DNADIST de la suite PHYLIP, un paquete libre para inferencia filogenética [34]. El programa DNADIST tiene distintas opciones para el cálculo de distancia. Para ambos conjuntos de muestra se usó la expresión de Jukes-Cantor dada por (3.4.5) y se eligió la opción para volcar el porcentaje de similaridad. La salida fue, en cada caso, la matriz de distancias de la muestra.

A partir de la matriz de distancias de cada muestra, expresada en porcentajes de similaridad, se procedió al agrupamiento en “clusters” observando que

las secuencias similares en un 95% o más, esto es con una disimilaridad menor o a lo sumo igual al 5%, correspondían a la misma OTU. Este umbral, en las situaciones más comunes y menos exigentes, permite discriminar las secuencias por especie. El software MOTHUR utiliza como criterio "default", para los agrupamientos jerárquicos de tipo aglomerativo que realiza, el del "vecino promedio". En [11] Everitt et al analizan los distintos métodos que pueden aplicarse. El método de encadenamiento individual, también llamado de "vecinos más cercanos", propone que la distancia entre dos grupos, u OTUs, sea la que hay entre sus miembros más próximos. Es decir; si U y V son dos grupos

$$d_{UV} = \min(d_{ij} : i \in U,\ j \in V) \tag{4.2.1}$$

En cambio el método de encadenamiento completo o de "vecinos más lejanos" establece la distancia entre dos grupos de acuerdo a

$$d_{UV} = \max(d_{ij} : i \in U,\ j \in V) \tag{4.2.2}$$

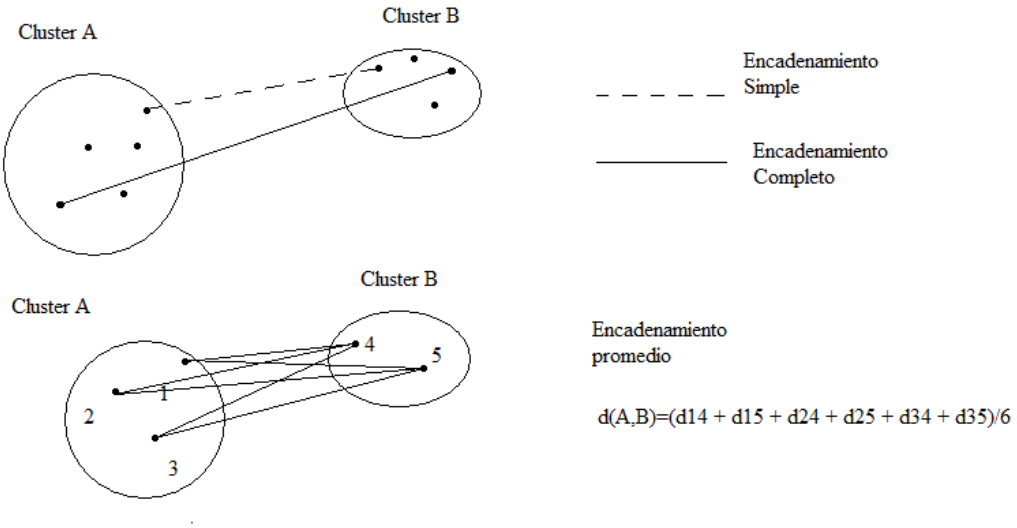
Esto implica que la distancia entre los dos grupos es la que hay entre sus miembros más alejados. El encadenamiento promedio consiste en tomar la distancia media entre todos los pares de secuencias, donde un elemento del par es de un grupo, y el otro elemento pertenece al otro grupo. Si n_U es el número de secuencias en U y n_V es el número de secuencias en V, entonces

$$d_{UV} = \frac{1}{n_U n_V} \sum_{i \in U} \sum_{j \in V} d_{ij} \tag{4.2.3}$$

En la Figura 4.2.1 cada secuencia es representada por un punto y se ven las tres opciones con las que se puede considerar la distancia entre "clusters".

Figura 4.2.1

Criterios de Agrupamiento



El procedimiento de cálculo parte de considerar inicialmente cada secuencia como un grupo unitario, es decir una OTU con un solo individuo, y las va juntando al estructurar un árbol de acuerdo a los niveles de disimilaridad que van creciendo desde las hojas hacia la raíz. Cuando se alcanza el umbral de disimilaridad establecido se "corta" el árbol y al nivel de ése "corte" cada rama representa una OTU distinta. La salida del proceso de agrupamiento realizado por MOTHUR es una planilla en la cual se listan cada una de las OTUs formadas con sus respectivas secuencias integrantes.

Para realizar todos estos procesos previos se tomó el criterio general de usar las opciones más comunes y de mayor simplicidad, pues se buscaba esencialmente analizar el desempeño de las estimaciones de riqueza una vez realizados los mismos. Sin embargo cabe mencionar que el reciente trabajo de Shloss [29] sugiere tener en cuenta la relación que existe entre el patrón de alineamiento utilizado y el nivel de "corte" del árbol filogenético obtenido. En todo caso, esta es entre otras, una de las razones, más ligada a consideraciones biológicas que estadísticas, que afecta a las mediciones de riqueza.

4.3 EVALUACION DE RIQUEZA Y DIVERSIDAD

Con MOTHUR se realizaron a continuación las estimaciones habituales de riqueza y diversidad. Para cada muestra de los dos conjuntos elegidos se establecieron las cantidades de especies observadas por suma total de los agrupamientos en OTUs con disimilaridad del 5%. También se calcularon las estimaciones de riqueza CHAO y ACE citadas en 3.2. Con la función de cálculo desarrollada en lenguaje R, adjunta en el CD anexo dentro del Script_2E.R, se evaluaron las respectivas entropías de las muestras. Además, se calcularon con la fórmula (3.2.7) las cantidades de especies que correspondería que existiesen, si sus distribuciones en cada muestra fueran uniformes. Los resultados para el conjunto SRX008158 se exhiben en la Tabla 4.3.1

Tabla 4.3.1

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
Cantidad de Secuencias	1641	8361	6926	6146	6226	8444	6103	5885
Entropía	5.938	7.759	7.049	6.852	6.767	6.669	6.687	6.878
N° Especies si hay uniformidad	379	2343	1152	946	869	788	802	971
Sobs	541	3575	2273	2030	1715	2040	1659	1842
CHAO	834	6351	4080	3363	2502	3180	2691	3110
ACE	1102	9525	5964	3755	2755	3544	3757	4048

El análisis de los datos muestra en primer lugar que ninguna de las muestras corresponde a comunidades microorgánicas uniformes pues las riquezas observadas, que denotadas por Sobs representan el total de especies que se cuenta en la muestra, y las estimadas superan sensiblemente a las calculadas bajo el supuesto de uniformidad del medio.

En segundo término surge que la cantidad observada de especies resulta aumentada por la estimación de CHAO que de todas formas es, según se cita en [12], sólo una cota inferior para la cantidad real de especies presentes en la comunidad. Por otro lado los valores que arroja la estimación ACE son sensiblemente mayores aunque también subestimen, como se comentó en 3.5, el total de especies existente en el medio.

Los valores obtenidos para las muestras del segundo conjunto ERX009564 se presentan en la Tabla 4.3.2

Tabla 4.3.2

Muestra	Err19	Err20	Err21	Err22	Err23	Err24
Cantidad de Secuencias	5011	5582	7637	3299	10371	5840
Entropía	6.132	6.212	6.825	6.431	7.200	6.613
N° Especies si hay uniformidad	460	499	921	621	1339	745
Sobs	957	1066	1930	997	2664	1358
CHAO	1821	1786	3374	1687	4937	2104
ACE	2239	2458	4531	2125	6237	2129

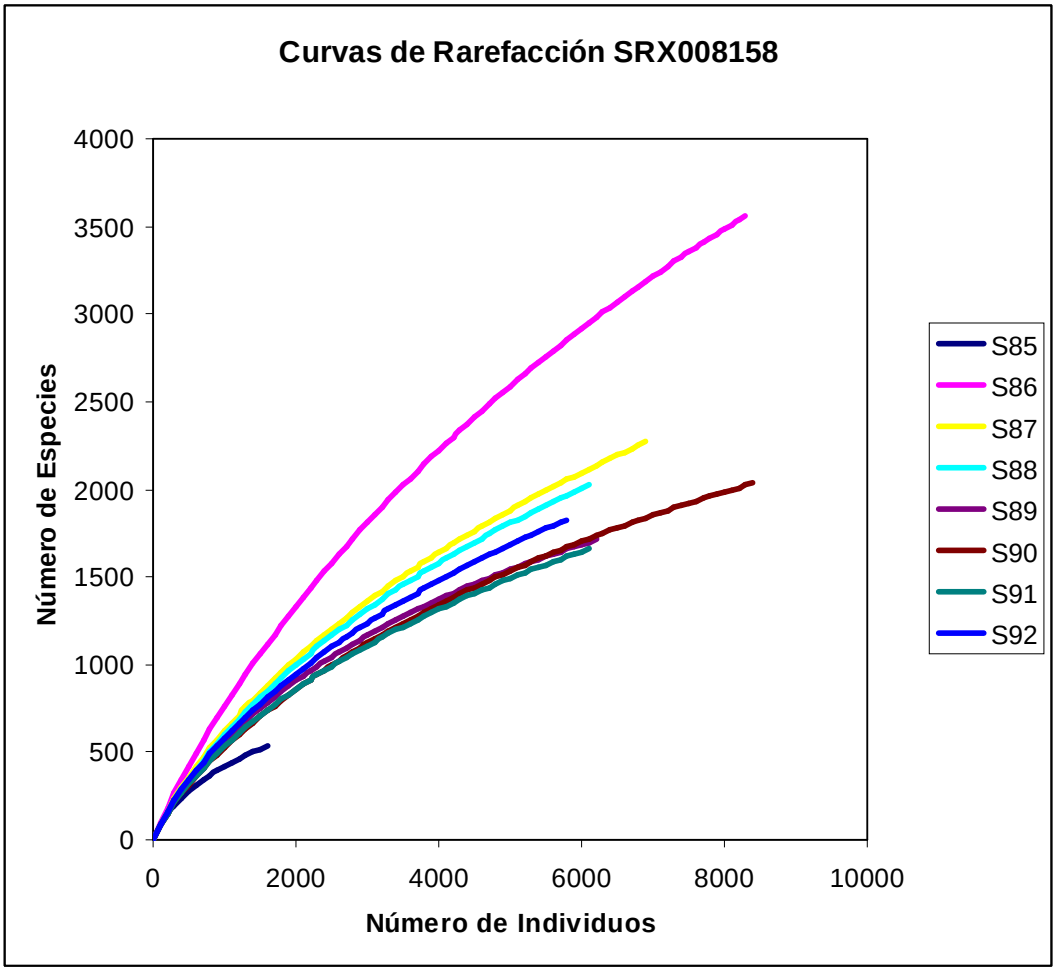
Aquí también el número de especies observadas supera ampliamente a las que corresponderían a muestras de distribución perfectamente uniformes. En cuanto a los estimadores CHAO y ACE se aprecian incrementos por sobre las cantidades de especies observadas que son, en términos relativos, similares a los obtenidos para el primer conjunto de muestras.

En términos generales al comparar las cantidades observadas de especies con las que debieran estar presentes si las muestras fueran uniformes, se evalúa la incertidumbre sobre la forma de una distribución poblacional que evidencia contener algunas especies en mayor cantidad que otras. Así se logra un primer paso en la aclaración del significado de los valores de entropía en el contexto de las evaluaciones de biodiversidad y se intenta despejar la objeción sobre la "*dificultad de entender lo que este estimador significa*" citada por Hill et al en [14].

Sobre las muestras de ambos conjuntos se realizó también un análisis de rarefacción que arrojó, para el primer conjunto considerado, las curvas exhibidas en la Figura 4.3.3

Figura 4.3.3

Análisis de Rarefacción del Conjunto de Muestras SRX008158



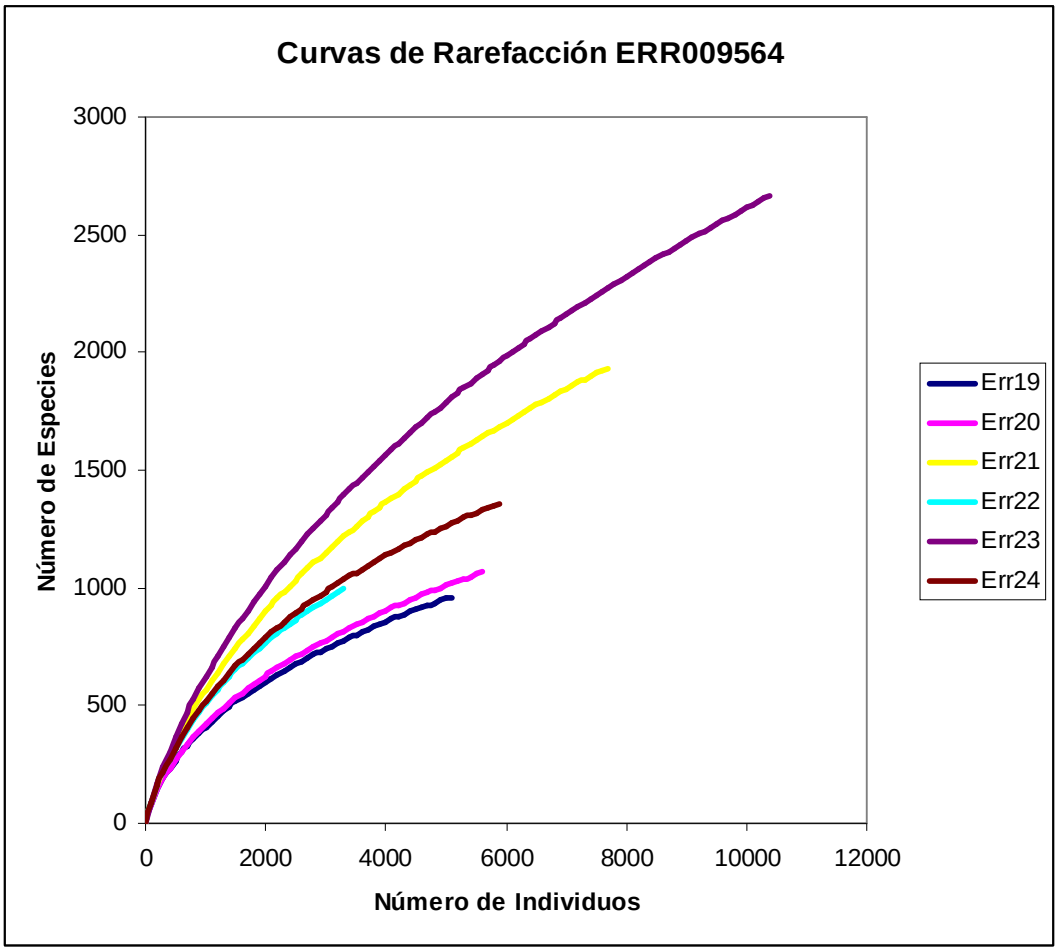
Los cálculos respectivos se hicieron utilizando MOTHUR que, por "default", realiza 1000 remuestreos con reposición para cada incremento de 100 individuos en la muestra. El número de especies en cada punto resulta entonces el promedio obtenido en esos 1000 remuestreos y no surge de ningún ajuste o cálculo teórico, razón por la cual la curva se interrumpe cuando se han considerado todos los individuos presentes en la muestra. Por esto, el número de especies resultante no es otro que el de las especies realmente observadas. Sin embargo esas cantidades adquieren diferente significación comparativamente pues, por ejemplo, al observar la figura surge claramente que la curvatura de las curvas S89, S90 y S91 para 6000 individuos está mucho mas cercana a alcanzar el comportamiento asintótico horizontal que la de la curva S86. En general puede observarse que el patrón gráfico en tal sentido varía

haciendo crecer la cantidad necesaria de individuos para alcanzar la asíntota horizontal según sean mayores los valores de entropía de las muestras. Una vez más se aprecia con mayor evidencia el significado de la entropía que aparecía difuso en [14].

Para el segundo conjunto de muestras, las curvas obtenidas son las de la Figura 4.3.4

Figura 4.3.4

Análisis de Rarefacción del Conjunto de Muestras ERR009564



En forma similar a la ya apuntada se observa, por ejemplo, que las curvas correspondientes al las muestras Err19 y Err20 considerados 4000 individuos se encuentran mas próximas a alcanzar la asíntota horizontal que la correspondiente a la muestra Err23. Esto a su vez se condice con los valores diferentes de entropía hallados que son bastante parecidos para Err19 y Err20 y sensiblemente mayores para Err23.

5-SIMULACIÓN

5.1 EL MODELO EXPERIMENTAL

Una idea para experimentar fue concebida a partir de dos conceptos. Por un lado se consideró la existencia de una estimación de la probabilidad de hallar una especie nueva cuando se selecciona un nuevo individuo para integrar la muestra. Por otro, se utilizó el hecho de que la curva de rarefacción debía alcanzar un comportamiento asintótico horizontal para un tamaño de la muestra suficientemente grande. La forma de estimación fue dada por Turing primero y luego desarrollada por Good en [21]. La fórmula (3.2.9) da cuenta de ella. En cuanto a la curva de rarefacción si ésta alcanza un comportamiento asintótico horizontal, esto será porque todas las curvas de acumulación a partir de las cuales pueda construirse, deberán observar en más o en menos un comportamiento similar cuando aumente el tamaño muestral. Sobre esta base conjunta se desarrolló el modelo experimental.

Cuando se trata de estimar la cantidad de especies presentes en un medio, al agregar un nuevo individuo éste puede resultar perteneciente a una especie ya conocida o no. La sucesión de los valores de cantidad de especies resulta entonces un proceso aleatorio como se describe a continuación. En el Diagrama 5.1.1 la variable i representa el número de individuos considerado en la muestra y S_i es la cantidad de especies halladas cuando la muestra tiene tamaño i.

Diagrama 5.1.1

Proceso Aleatorio de Cantidad de Especies

…	S_{i-2}	S_{i-1}	S_i	S_{i+1}	S_{i+2}	…
…	i-2	i-1	i	i+1	i+2	…

Para estimar la probabilidad de que el i-ésimo individuo agregado corresponda a una especie nueva se toma el estimador de Turing

$$\hat{T}_i = \frac{n°sgletones}{i-1} \qquad (5.1.1)$$

donde cada singletón en la muestra de i-1 individuos está formado por el solo individuo que representa a una especie en esa muestra. Resulta entonces que para cada valor de S_i hay una probabilidad asociada como se ve en la Tabla 5.1.2

Tabla 5.1.2

Estado	Probabilidad estimada
$S_i=S_{i-1}$	$p = 1 - \hat{T}_i$
$S_i=S_{i-1}+1$	$p = \hat{T}_i$

Así el valor esperado de S_i, que correspondería a la curva de rarefacción para i individuos considerados se calcula:

$$E(S_i) = S_{i-1}(1-\hat{T}_i) + (S_{i.-1}+1)\hat{T}_i \qquad (5.1.2)$$

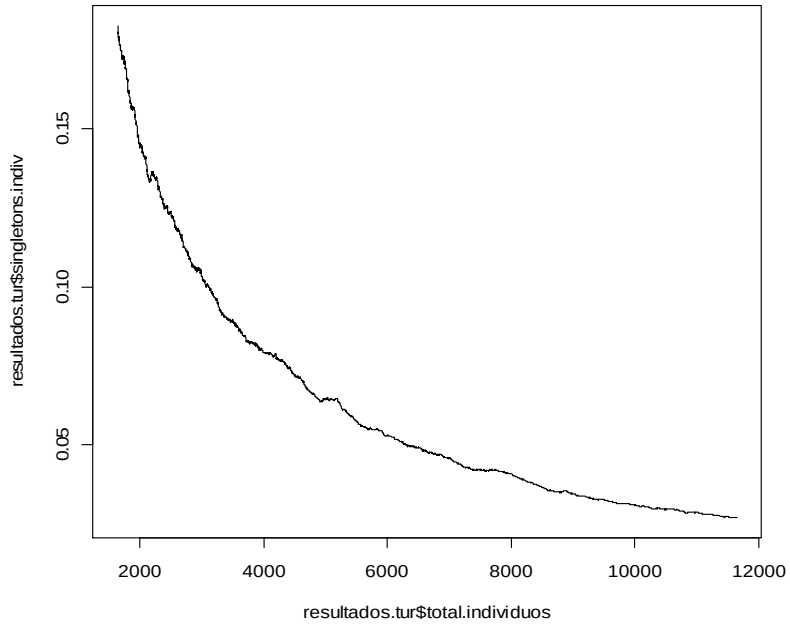
y operando se obtiene

$$E(S_i) = S_{i-1} + \hat{T}_i \qquad (5.1.3)$$

Si se realiza una simulación eligiendo de a uno individuos en una muestra, cabría esperar que cuando i crezca $\hat{T}_i \to 0$ pues el número de especies aún no encontradas debiera ir disminuyendo al ser finita la cantidad S de especies buscada. En la Figura 5.1.3 se aprecia como efectivamente se reduce el cociente $\hat{T}_i$ de la muestra, conforme aumenta el número de individuos que la integra.

Figura 5.1.3

Disminución de la Probabilidad de Especie Nueva



En esta situación, también debiera ocurrir entonces que $E(S_i) \to S$ al crecer el tamaño i.

5.2 ALGORITMO DE RECUENTO DE ESPECIES (ARE)

- *La Secuencia de Cálculo*

El algoritmo realiza la simulación por la técnica de Monte Carlo. Dado el tamaño n de la muestra original se determina el valor del estimador de la probabilidad de especie nueva. Ese valor permite constituir los intervalos $[0, \hat{T}_n]$ y $(\hat{T}_n, 1]$ de modo que al elegir un número aleatorio r tal que $0 \le r \le 1$, si cae dentro del primer intervalo el nuevo individuo simulado corresponda a una especie nueva y si cae dentro del segundo intervalo sea un ejemplar de una especie conocida. Si ocurre lo primero, la cantidad de especies en el medio se incrementa en 1 y si no, se utilizan las proporciones existentes de cada especie para asignar por medio de un nuevo número aleatorio la especie ya conocida, a la cual pertenece el nuevo individuo. Así se van agregando individuos hasta que la cuenta de las especies nuevas alcance un valor estable o hasta que se cumpla con

algún otro criterio de corte de la simulación. Los pasos del procedimiento se sintetizan en forma secuenciada a continuación.

1- Dada la muestra elegida, de tamaño n, y su agrupamiento en OTUs, se determina el valor inicial del estimador de Turing $\hat{T}_{i+1} = \frac{f_1}{i}$ siendo $i = n$

2- Se elige un número aleatorio r, tal que $0 \le r \le 1$ y se pregunta si está en el intervalo $[0, \hat{T}_{i+1}]$. Si es así, se realiza $S_{i+1} = S_i + 1$ y se va al paso 4. Si ocurre lo contrario se realiza $S_{i+1} = S_i$ y se va al paso 3

3- Se utiliza la distribución de abundancia de la muestra (ver Figura 2.4.1) para calcular la proporción de individuos que están en OTUs de $1,2,...n$ individuos y con estas proporciones se determina, por un sorteo de acuerdo a ellas, a que grupo de OTUs ya conocidas pertenece el nuevo individuo. Para establecer a que OTU específica, de entre las de este grupo, corresponde el nuevo individuo se realiza un nuevo sorteo con probabilidad uniforme para cada OTU del grupo.

4- Sea el nuevo individuo de una nueva especie o no, la muestra tiene ahora un elemento más. Se pregunta entonces si el procedimiento debe cortarse porque se cumple el criterio elegido para ello en cuyo caso la simulación ha finalizado. Si el criterio de corte no se cumple, se asigna entonces $i \leftarrow i+1$, se calcula la nueva distribución de abundancia y la nueva estimación de Turing según $\hat{T}_{i+1} = \frac{f_1}{i}$ y se repite desde el paso 2.

El programa de computadora correspondiente fue desarrollado en lenguaje R y se presenta en el CD anexo bajo el nombre Pruebas_2.R

- *Resultados*

Como primera alternativa para fijar el corte del procedimiento resultó natural hacerlo al alcanzar un número prefijado de individuos simulados. De acuerdo a ello se procesaron las muestras del conjunto SRX008158 que arrojaron los resultados de la Tabla 5.2.1

Tabla 5.2.1

Muestra	N°Individuos	Sobs	CHAO	ACE	ARE	Número de Individuos. Simulados
S85	1641	541	834	1102	997	4000
S86	8361	3575	6351	9595	6650	15000
S87	6926	2273	4080	5964	4271	14000
S88	6146	2030	3363	3755	3920	13000
S89	6226	1715	2502	2755	3067	13000
S90	8444	2040	3180	3544	3537	15000
S91	6103	1659	2691	3757	3065	13000
S92	5885	1842	3110	4048	3417	12000

De la observación de los resultados surge claramente que la simulación efectuada arroja valores de cantidad de especies similares o apenas un poco superiores a los estimados por CHAO, que constituye una cota inferior. Por otra parte las cantidades obtenidas son sensiblemente menores que las estimadas por ACE, las que también suelen subestimar la verdadera riqueza de especies. Estos resultados sugirieron entonces trabajar en dos direcciones. Por un lado aumentando el tamaño de la muestra simulada, lo que sin duda traería aparejado mayores tiempos de proceso y, por otro, intentando mejorar el estimador. Antes de realizar pruebas incrementando el tamaño muestral se procuró ajustar el estimador, adecuándolo un poco más a la incertidumbre contenida en el problema.

5.3 ALGORITMO DE RECUENTO DE ESPECIES CON COBERTURA (AREC)

- *Estimación de la Cobertura*

En [20] Chao y Lee presentan la idea de cobertura que aplican para derivar el estimador conocido como ACE. Más tarde, Chao y Bunge utilizan la misma idea en [37] para construir un estimador del número de especies a partir de modelos de abundancia de tipo paramétrico no pensados, en principio, para comunidades microbianas como las que se analizan.

Para el modelo multinomial definido en 3.1 y aquí utilizado, cada una de las S especies existentes en el medio tiene una probabilidad p_j de aparición. Si se toma

una muestra de tamaño i, de forma tal que a cada especie le correspondan x_j individuos de la misma, se define la cobertura como

$$C = \sum_{j=1}^{S} p_j I[x_j > 0] \tag{5.3.1}$$

dónde I es la función indicador que vale 1 si $x_j > 0$ y 0 en otro caso. El valor de S es desconocido y, en realidad, en la expresión de C solo suman aquellas especies que efectivamente aparecen. Claramente $0 \le C \le 1$. Si $C = 0$ es porque no ha aparecido aún ninguna especie (caso solo teórico e imposible si se tomó una muestra) y si $C = 1$ es porque todas las especies existentes han aparecido en la muestra.

Además a partir de la muestra puede calcularse el número de especies representadas por r individuos

$$f_r = \sum_{j=1}^{S} I[x_j = r] \tag{5.3.2}$$

suponiendo una cantidad f_0 que sea precisamente el número de especies con 0 individuos. Obsérvese que puede ocurrir que $f_r = 0$ para varios valores de $r = 1,...,i$. Así

$$\sum_{r=1}^{i} f_r = S_i \tag{5.3.3}^{*}$$

dónde S_i es la cantidad de especies halladas en la muestra que tiene i individuos y claramente resulta

$$S = S_i + f_0 \tag{5.3.4}$$

Además f_1 es el número de singletones en la muestra y $\sum_{r=1}^{i} r f_r = i$ es el tamaño muestral.

Según exponen Chao y Shen [17], un estimador de la cobertura según la muestra tomada es

$$\hat{C} = 1 - \frac{f_1}{i} = 1 - \hat{T}_i \tag{5.3.5}$$

---

* Para adaptarse al concepto algorítmico, se varía la notación respecto de la utilizada en (3.2.3) y (3.2.4). Ahora $D = S_i$ y $N = S$

y la probabilidad p_j de elección de un individuo de la j-ésima especie se estima por

$$\hat{p}_j = \frac{x_j}{i}\hat{C} \qquad (5.3.6)$$

- *La Secuencia de Cálculo*

En el presente trabajo se utiliza la idea de cobertura para corregir la estimación de Turing de la fórmula (5.1.1) en el paso i de una simulación que agrega individuos a la muestra para estimar la cantidad S. Así, en la medida que se va incrementando el tamaño muestral y aparecen nuevas especies, va creciendo también la cobertura y, simultáneamente ocurre que $\hat{T}_i \to 0$. El procedimiento de simulación es el que sigue.

1- Dada la muestra elegida, de tamaño n, y su agrupamiento en OTUs, se determina el valor inicial del estimador de Turing $\hat{T}_{i+1} = \frac{f_1}{i}$ siendo $i = n$.

2- Se calcula la estimación de cobertura mediante $\hat{C} = 1 - \frac{f_1}{i} = 1 - \hat{T}_{i+1}$

3- Como la muestra actual tiene una cobertura estimada $\hat{C}$ y cada especie que resultó un singletón tiene una frecuencia relativa en la muestra dada por $\frac{1}{i}$, esta probabilidad puede ser corregida por la cobertura estimada de la muestra de modo que la probabilidad de cada singleton resultará $\frac{1}{i}(1 - \frac{f_1}{i})$. Como esto debe ocurrir para los f_1 singletones hallados se obtiene una probabilidad de especie nueva corregida $p_{ns} = \hat{T}_{i+1}\hat{C} = \frac{f_1}{i}(1 - \frac{f_1}{i})$

4- Se elige un número aleatorio r, tal que $0 \le r \le 1$ y se pregunta si está en el intervalo $[0, p_{ns}]$. Si es así, se realiza $S_{i+1} = S_i + 1$ y se va al paso 6. Si ocurre lo contrario se realiza $S_{i+1} = S_i$ y se va al paso 5

5- Se utiliza la distribución de abundancia de la muestra, sin corrección por cobertura, para calcular la proporción de individuos que están en OTUs de 1,2,...i individuos. Con estas proporciones se determina, por un sorteo de

acuerdo a ellas, a que grupo de OTUs ya conocidas pertenece el nuevo individuo. Para establecer a que OTU específica, de entre las de este grupo, corresponde el nuevo individuo se realiza un nuevo sorteo con probabilidad uniforme para cada OTU del grupo.

6- Sea el nuevo individuo de una nueva especie o no, la muestra tiene ahora un elemento más. Se pregunta entonces si el procedimiento debe cortarse porque se cumple el criterio elegido para ello, en cuyo caso la simulación ha finalizado. Si el criterio de corte no se cumple, se asigna entonces $i \leftarrow i+1$, se calcula la nueva distribución de abundancia, la nueva estimación de Turing según $\hat{T}_{i+1} = \frac{f_1}{i}$ y se repite desde el paso 2.

El algoritmo fue programado en lenguaje R e incorporado al CD anexo con el nombre Pruebas2.R

Hay que observar que en cada paso iterativo resulta $p_{ns} \leq \hat{T}_{i+1}$ de tal modo que este ajuste de la estimación de Turing a la incertidumbre que parte de la muestra deberá redundar en una disminución del valor de la riqueza estimada, a igual número de iteraciones, respecto de la obtenida sin corrección por cobertura.

- *Resultados*

En la Tabla 5.3.1 se muestran los resultados obtenidos para el conjunto de muestras SRX008158

Tabla 5.3.1

Muestra	N° Individuos	Sobs	CHAO	ACE	ARE	AREC	Número de Individuos. Simulados	Cobertura Final
S85	1641	541	834	1102	997	938	4000	0.94698
S86	8361	3575	6351	9595	6650	5754	15000	0.90765
S87	6926	2273	4080	5964	4271	3834	14000	0.93347
S88	6146	2030	3363	3755	3920	3552	13000	0.93204
S89	6226	1715	2502	2755	3067	2882	13000	0.95292
S90	8444	2040	3180	3544	3537	3307	15000	0.95184
S91	6103	1659	2691	3757	3065	2928	13000	0.94607
S92	5885	1842	3110	4048	3417	3163	12000	0.93743

Como se esperaba, a igual número de iteraciones, la simulación sin corrección por cobertura da valores de riqueza levemente superiores a los obtenidos simulando con cobertura. Si embargo debe observarse que la cobertura del espacio de probabilidad dada en cada muestra simulada estuvo en el orden del 95% del total. Hay que resaltar entonces que en el 5% restante y aún no visible podrían estar las especies "raras" de baja probabilidad.

5.4 SUAVIZACIONES (ARECS1 Y ARECS2)

- *Fórmulas de Suavizado*

A fin de obtener un mayor ajuste a la real incertidumbre planteada por el tamaño escaso de la muestra frente al de la población y a la diversidad no uniforme de la distribución de las especies, se desarrollaron dos alternativas de simulación adicionales para "suavizar" incluso el efecto de la cobertura.

ARECS1: En el paso actual i, se calcula la cobertura corregida como el promedio de la actual y la anterior: $\tilde{C}_i = \frac{\hat{C}_{i-1} + \hat{C}_i}{2}$ para $i \succ n$ y como inicialmente $i = n$ se toma, en ese caso, $\hat{C}_{n-1} = \hat{C}_n$ En cada iteración se calcula entonces la probabilidad de especie nueva $p_{ns} = \hat{T}_{i+1}\tilde{C}$ y se siguen los pasos indicados en 5.3

ARECS2: En el paso actual i de la simulación se utiliza un promedio de todas las coberturas hasta ahora calculadas: $\overline{C}_i = \frac{\sum_{j=n}^{i} \hat{C}_j}{i-n+1}$ donde $\hat{C}_j = 1 - \hat{T}_{j+1}$. Es decir, en cada iteración habrá una cobertura de valor $\hat{C}_i$ y otra suavizada $\overline{C}_i$. Con esta última se calcula el valor de la probabilidad de especie nueva p_{ns} y se sigue el flujo del algoritmo detallado en 5.3

Los procedimientos de cálculo relativos a estos suavizados fueron programados en lenguaje R e incorporados en el CD anexo con el nombre Pruebas2.R

- Resultados

Las pruebas sobre el conjunto de muestras SRX008158, realizadas para iguales cantidades de individuos simulados, arrojaron los valores que se dan en la Tabla 5.4.1

Tabla 5.4.1

Muestra	N° Individuos	Sobs	CHAO	ACE	ARE	ARES	ARECS1	ARECS2	Número de Individuos. Simulados
S85	1641	541	834	1102	997	938	956	935	4000
S86	8361	3575	6351	9595	6650	5754	5899	5740	15000
S87	6926	2273	4080	5964	4271	3834	3974	3806	14000
S88	6146	2030	3363	3755	3920	3552	3550	3519	13000
S89	6226	1715	2502	2755	3067	2882	2873	2825	13000
S90	8444	2040	3180	3544	3537	3307	3356	3281	15000
S91	6103	1659	2691	3757	3065	2928	2953	2775	13000
S92	5885	1842	3110	4048	3417	3163	3285	3197	12000

Se observa que ARECS1 1 logra en la mayoría de las muestras elevar someramente las estimaciones de riqueza realizadas con cobertura. Para ARECS2 ocurre lo inverso y las estimaciones resultan apenas inferiores que las hechas con cobertura. La estimación de ambos suavizados permanece siempre inferior a la calculada por ARE.

Con las distintas opciones de estimación desarrolladas se realizó una prueba sobre el conjunto de muestras ERX009564 para confirmar el desempeño hasta aquí observado. Los resultados se exponen en la Tabla 5.4.2

Tabla 5.4.2

Muestra	N° Individuos	Sobs	CHAO	ACE	ARE	AREC	Cobertura Final	ARECS1	ARECS2	Número de Individuos Simulados
Err19	5011	957	1821	2239	1735	1639	0.96508	1677	1670	10000
Err20	5582	1066	1786	2458	1866	1846	0.96309	1785	1808	11000
Err21	7637	1930	3374	4531	3508	3229	0.95224	3356	3130	15000
Err22	3299	997	1687	2125	1881	1664	0.94765	1692	1691	7000
Err23	10371	2664	4937	6237	4942	4545	0.94774	4663	4515	20000
Err24	5860	1358	2104	2129	2346	2256	0.95711	2251	2221	11000

Los valores obtenidos replican lo ocurrido con el primer conjunto de muestras procesado. Para la cantidad de individuos simulados, como en el primer conjunto aproximadamente el doble que los individuos reales de la muestra, el estimador de Turing arrojó valores de riqueza levemente superiores a los calculados por CHAO y bastante menores que los estimados por ACE. La simulación con cobertura dio valores ligeramente inferiores a la de Turing, alcanzando una cobertura de alrededor del 95% del espacio de probabilidad. El Suavizado 1 obtuvo estimaciones en general levemente superiores a las de cobertura pero inferiores a las de Turing. El Suavizado 2 arrojó valores alrededor de los obtenidos con cobertura, en la mitad de los casos muy ligeramente por arriba de los mismos y en la otra mitad muy ligeramente por debajo.

5.5 PRUEBAS

De acuerdo a las direcciones de trabajo establecidas en 5.2 y exploradas ya algunas nuevas posibilidades de estimación, se decidió utilizar, en principio, ARE y AREC aumentando el tamaño de las muestras simuladas. ARE aparecía hasta aquí como la estimación que arrojaba más altos valores de riqueza, aunque esto resultara inferior al desempeño de ACE. Y AREC tenía un comportamiento intermedio entre las leves variaciones que frente a él ofrecían los dos algoritmos con suavizados. Se decidió también incorporar el cálculo de la entropía “de partida”, es decir de la entropía de las muestras de individuos reales medida antes de la simulación. Este agregado se realizó a efecto de analizar el vínculo entre riqueza estimada y diversidad inicial detectada en las muestras.

- *Prueba 1*

La primera prueba consistió en triplicar la cantidad de individuos simulados. Para el conjunto de muestras SRX008158 se obtuvieron los resultados detallado en la Tabla 5.5.1

Tabla 5.1.1

Muestra	N° Individuos	Sobs	CHAO	ACE	ARE	AREC	Cobertura Final	Entropía Inicial	Número de Individuos Simulados
S85	1641	541	834	1102	1261	1165	0.97991	5.938	12000
S86	8361	3575	6351	9595	9101	7682	0.95532	7.759	45000
S87	6926	2273	4080	5964	5472	5140	0.97118	7.049	42000
S88	6146	2030	3363	3755	5467	4645	0.97182	6.852	39000
S89	6226	1715	2502	2755	3813	3583	0.98051	6.767	39000
S90	8444	2040	3180	3544	4768	4113	0.97973	6.669	45000
S91	6103	1659	2691	3757	4088	3646	0.97893	6.687	39000
S92	5885	1842	3110	4048	4597	4144	0.97192	6.878	36000

Se observa que la estimación de ARE resulta sensiblemente superior a la de CHAO y en la mayoría de los casos superior también a la de ACE. En las muestras S86 y S87 el valor inferior de ARE está asociado a los más altos valores de la entropía "de partida", lo que sugiere que aumentando el número de individuos simulados, también para estas muestras, ARE podría alcanzar un mejor desempeño que ACE. La simulación con cobertura AREC ha producido estimaciones sensiblemente superiores a CHAO, lo que es acorde con el crecimiento obtenido en el porcentaje de cobertura del espacio de probabilidad que, en general, supera ahora el 97%. Esto no ocurre justamente para la muestra S86, la cual registra la máxima entropía "de partida" del conjunto. Así se refuerza la idea de que el aumento de la cantidad de individuos simulados redundará en mayores valores de riqueza estimada.

Con las muestras del conjunto ERX009564 se buscó confirmar estos comportamientos realizando una simulación en condiciones similares. Los valores calculados se detallan en la Tabla 5.5.2

Tabla 5.5.2

Muestra	N° Individuos	Sobs	CHAO	ACE	ARE	AREC	Cobertura Alcanzada	Entropía Inicial	Número de Individuos Simulados
Err19	5011	957	1821	2239	2134	2262	0.98186	5.922	30000
Err20	5582	1066	1786	2458	2500	2348	0.98364	5.979	33000
Err21	7637	1930	3374	4531	4595	4152	0.97900	6.640	45000
Err22	3299	997	1687	2125	2385	2203	0.97460	6.260	21000
Err23	10371	2664	4937	6237	6409	5985	0.97693	6.974	60000
Err24	5860	1358	2104	2129	2844	2768	0.98203	6.391	33000

El análisis de los resultados confirma en líneas generales comportamientos de los estimadores similares a los observados para el otro conjunto de muestras. La simulación ARE arrojó cantidades de especies casi siempre superiores a las estimadas por ACE, salvo para la muestra Err19 en que resultó menor. Este fue precisamente el caso con entropía inicial menor, en el cual además la simulación por cobertura AREC también arrojó un valor superior a la ARE. Estos hechos sugieren que en la medida en que la entropía inicial disminuye, la simulación por cobertura AREC obtiene estimaciones cercanas a la simulación ARE o aún mayores. En forma complementaria se observa también que en los resultados sobre ambos conjuntos de muestras las diferencias relativas entre las simulaciones ARE y con cobertura AREC, para una misma muestra, se amplían a medida que crece el valor de la entropía inicial.

- *Prueba 2*

En la segunda prueba se realizó la simulación utilizando un criterio de corte fuertemente sugerido por la prueba anterior. Esto consistió en simular el agregado de individuos hasta que la cobertura alcanzara un cierto valor mínimo, en vez de establecer un número fijo de individuos a simular. En esta oportunidad se trabajó sobre el conjunto de muestras ERX009564 y los resultados obtenidos se exhiben en la Tabla 5.5.3

Tabla 5.5.3

Muestra	AREC (N° Indvs. Fijo)	Cobertura Alcanzada	Número de Individuos Simulados	AREC con Corte	Cobertura de Corte	Número de Individuos. Simulados	Entropía Inicial
Err19	2262	0.98186	30000	1768	0.975	15272	5.922
Err20	2348	0.98364	33000	1885	0.975	15852	5.979
Err21	4152	0.97900	45000	4053	0.975	37806	6.640
Err22	2203	0.97460	21000	2493	0.975	24312	6.260
Err23	5985	0.97693	60000	5769	0.975	65419	6.974
Err24	2768	0.98203	33000	2840	0.975	32142	6.391

Las estimaciones obtenidas al cortar la simulación cuando la cobertura alcanza el valor 0.975 confirman, en general, la idea de vincular la necesidad de simular más casos para determinar la riqueza, con el crecimiento de las entropías iniciales. Las

muestras Err19 y Err20, que son las de menor entropía, alcanzan con la mitad de casos que la Err21, por ejemplo, la cobertura de corte propuesta. En cambio la muestra Err23, que es la de mayor entropía, requirió más casos que los fijados en la primera prueba para alcanzar el nivel de cobertura de corte. Sin embargo, aquí hay que observar que con los 60000 casos de la simulación con cobertura AREC, a número fijo de individuos, se logró una estimación mayor, y también una cobertura mayor, que cuando en la presente prueba se simularon 65419 individuos. Esto sugiere que una sola simulación para cada muestra puede no ser suficiente y que sería conveniente establecer un número de simulaciones a efecto de hallar un intervalo de confianza para la estimación media.

Sobre el conjunto SRX008158 de muestras se realizó la misma prueba cuyos resultados se exponen en la Tabla 5.5.4

Tabla 5.5.4

Muestra	AREC (N° Indvs Fijo)	Cobertura Alcanzada	Número de Individuos. Simulados	AREC con Corte	Cobertura de Corte	Número de Individuos Simulados	Entropía Inicial
S85	1165	0.97991	12000	1111	0.975	9496	5.938
S86	7682	0.95532	45000	8003	0.969	60000	7.759
S87	5140	0.97118	42000	5535	0.975	54249	7.049
S88	4645	0.97182	39000	4893	0.975	47896	6.852
S89	3583	0.98051	39000	3452	0.975	37282	6.767
S90	4113	0.97973	45000	4226	0.975	49882	6.669
S91	3646	0.97893	39000	3497	0.975	32420	6.687
S92	4144	0.97192	36000	4088	0.975	37526	6.878

Los hechos observados vuelven a evidenciar, en general, la relación existente entre la estimación del valor de la riqueza y el número de individuos simulados usando la cobertura. Para las muestras S87 y S88, en las cuales el corte de cobertura a 0.975 resultó mayor que el valor de cobertura obtenido con la simulación de cantidad fija de individuos, las estimaciones resultaron levemente mayores. En las mismas condiciones, sin embargo, la simulación con la muestra S92 arrojó una estimación menor en la simulación con cobertura y corte. Como en estas tres muestras mencionadas la entropía es bastante parecida, se asoció este efecto al particular caso aleatorio simulado y se concluyó en la conveniencia de reiterar las simulaciones para obtener un promedio como estimación de riqueza y un intervalo de confianza para el

valor estimado. La muestra S86, la de más alta entropía inicial, no alcanzó a generar un valor de cobertura mayor o igual a 0.975 en las 60000 simulaciones colocadas como límite al “loop” de programación, lo que indica que para entropías iniciales altas se requiere mayor cantidad de individuos simulados para alcanzar mayor cobertura y lograr así mayor precisión en la estimación de riqueza.

- *Prueba 3*

La tercera prueba se realizó al combinar los resultados y sugerencias de los análisis de las dos pruebas anteriores. Por un lado se estableció un corte distinto a la simulación ARE. En vez de considerar una cantidad fija de individuos simulados como en 5.2, se requirió simular casos hasta que la proporción de OTUs singletones, en el total de individuos en la muestra simulada, estuviera por debajo de un cierto umbral. Para el caso se estableció ese valor de corte en 0.03 Si la proporción de singletones, que es el estimador de Turing para la probabilidad de nueva especie dado en (5.1.1), es baja, esto implica que quedan pocas especies por descubrir. En ese marco, cada vez más agrupamientos u OTUs tienen más de un elemento, conforme avanza la simulación. En segundo lugar se decidió realizar cada simulación un número fijo de veces a efecto de establecer un intervalo de confianza para la estimación de riqueza. Esto tanto para ARE con corte según la proporción de singletones, como para AREC con corte según valor de cobertura. En tercer término se calculó, para ambos tipos de simulación, la entropía “de salida” o “de corte”, que es la entropía de la muestra cuando ya se han agregado todos los individuos simulados a la muestra. Una vez más, los respectivos programas fueron realizados en lenguaje R y se adjuntan en el CD anexo bajo los nombres Prueba_Corte_Turing.R y Prueba_Corte_Cobertura:R

Se trabajó primero sobre el conjunto de muestras SRX008158 con resultados que se expresan en la Tabla 5.5.5

Tabla 5.5.5

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
N° Individuos	1641	8361	6926	6146	6226	8444	6103	5885
Entropía Inicial	5.938	7.759	7.049	6.852	6.767	6.669	6.687	6.878
Sobs	541	3575	2273	2030	1715	2040	1659	1842
CHAO	834	6351	4080	3363	2502	3180	2691	3110
ACE	1102	9525	5964	3755	2755	3544	3757	4048
Número de Simulaciones	7	7	7	7	7	7	7	7
Confianza del Intervalo	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Media ARE/Corte=0.03	1288	9851	6015	5567	3759	4384	3924	4800
Intervalo de Confianza ARE	1215-1361	9542-10159	5706-6323	5381-5753	3646-3872	4237-4532	3816-4032	4542-5058
Cantidad Promedio de Individuos Simulados	10491	60000	51615	48292	29859	35728	32989	41156
Entropía Final Promedio	6.304	8.326	7.546	7.387	7.106	6.995	7.071	7.352
Media AREC/ Corte=0.975	1088	8165	5355	4743	3482	4223	3586	4286
Intervalo de Confianza AREC	997-1179	7999-8331	5218-5492	4601-4884	3302-3663	4072-4375	3461-3711	4137-4434
Cantidad de Individuos Simulados	9413	60000	51556	44973	31102	39888	33768	40900
Entropía Final Promedio	6.186	8.135	7.430	7.245	7.036	6.967	6.985	7.234

El análisis de los valores detallados permite establecer que la estimación promedio de ARE, utilizando una proporción de corte de 0.03, supera en todos los casos a la estimación respectiva realizada por ACE. Además salvo para la muestra S87, para todas las otras, la estimación ACE está por debajo del límite inferior del intervalo de confianza del 95% construido para la estimación de la riqueza de especies según este método. Por otra parte la estimación promedio realizada por AREC hasta cubrir el 97.5% del espacio de probabilidad, oscila en torno de los valores respectivos de ACE superándolos para las muestras S88, S89, S90 y S92 y quedando por debajo de ellos en el resto de los casos. Para todas las muestras, la estimación de CHAO queda sensiblemente por debajo del límite inferior del intervalo de confianza del 95% establecido para la estimación del número de especies, efectuada de acuerdo a este procedimiento. En el caso de la muestra S86, las cantidades promedio de individuos simulados por uno y otro método revelan que no fueron alcanzados los umbrales de

corte previstos en ambos procedimientos a pesar de lo cual las estimaciones estuvieron a tono con las realizadas para las otras muestras. Por último, la entropía final promedio, calculada sobre las muestras, destaca un crecimiento respecto de la exhibida por las muestras reales iniciales. Este incremento evidencia el aumento de la diversidad producido por el agregado de especies.

Con la finalidad de precisar y confirmar los análisis expuestos, se realizaron las simulaciones, en similares condiciones, para el conjunto de muestras ERX009564. Sus resultados son los de la Tabla 5.5.6

Tabla 5.5.6

Muestra	Err19	Err20	Err21	Err22	Err23	Err24
N° Individuos	5011	5582	7637	3299	10371	5840
Entropía Inicial	6.132	6.212	6.825	6.431	7.200	6.613
Sobs	957	1066	1930	997	2664	1358
CHAO	1821	1786	3374	1687	4937	2104
ACE	2239	2458	4531	2125	6237	2129
Número de Simulaciones	7	7	7	7	7	7
Confianza del Intervalo	0.95	0.95	0.95	0.95	0.95	0.95
Media ARE/Corte=0.03	1856	2102	4357	2338	6244	2732
Intervalo de Confianza ARE	1779-1933	2020-2185	4116-4597	2204-2473	6033-6454	2632-2831
Cantidad Promedio de Individuos Simulados	14199	16666	35664	19071	53136	21090
Entropía Final Promedio	6.364	6.461	7.186	6.790	7.566	6.866
Media AREC/Corte=0.975	1810	2035	4076	2167	5668	2579
Intervalo de Confianza AREC	1749-1872	1973-2098	3980-4173	2054-2280	5605-5730	2471-2686
Cantidad Promedio de Individuos Simulados	15856	18117	38152	19433	52253	22082
Entropía Final Promedio	6.341	6.429	7.112	6.724	7.493	6.833

Los valores exhibidos en la Tabla 5.5.6 indican que, para el conjunto de muestras analizado, la estimación ARE no presenta resultados tan concluyentes como en el caso anterior. En particular para las muestras Err19, Err20 y Err21 los valores de la estimación ACE son superiores a los obtenidos por el procedimiento de simulación, mientras que para las muestras Err22 y Err23 resultan prácticamente iguales. Solo en la

muestra Err24 la estimación ARE es mayor que la ACE. Además el límite superior del intervalo de confianza hallado para los casos Err19 y Err20 resulta menor que la estimación ACE y en los casos de las muestras Err21, Err22 y Err23 ésta cae dentro del intervalo. Todo esto sugiere la necesidad de variar el parámetro de corte exigiendo una proporción de singletones menor para cortar la simulación. Lo mismo vale para la estimación AREC que para las muestras Err19 y Err20 presenta una estimación CHAO que incluso cae dentro del intervalo de confianza construido. Nuevamente, corresponderá exigir una cobertura mayor para cortar la simulación. Si se analiza la entropía se ve que las muestras Err19 y Err20 son las que tienen entropía inicial menor. Esto quizás explique la necesidad de ajuste de los parámetros de corte, pues las especies no halladas podrían contribuir de manera más significativa a la diversidad que en otros casos. Respecto de la variación de la entropía de la inicial a la final, se confirma el comportamiento observado para la simulación del primer conjunto de muestras: el aumento de entropía se corresponde con el aumento de la diversidad por agregado de nuevas especies.

La simulación sobre el mismo conjunto de muestras con parámetros de corte más exigentes arrojó los resultados de la Tabla 5.5.7

Tabla 5.5.7

Muestra	Err19	Err20	Err21	Err22	Err23	Err24
N° Individuos	5011	5582	7637	3299	10371	5840
Entropía Inicial	6.132	6.212	6.825	6.431	7.200	6.613
Sobs	957	1066	1930	997	2664	1358
CHAO	1821	1786	3374	1687	4937	2104
ACE	2239	2458	4531	2125	6237	2129
Número de Simulaciones	7	7	7	7	7	7
Confianza del Intervalo	0.95	0.95	0.95	0.95	0.95	0.95
Media ARE/Corte=0.015	2291	2526	5244	2861	7124/0.182	3253
Intervalo de Confianza ARE	2170-2412	2441-2611	4969-5518	2758-2964	7005-7243	3116-3390
Cantidad Promedio de Individuos Simulados	33988	38431	79357	43157	90000	47389
Entropía Final Promedio	6.427	6.497	7.218	6.844	7.608	6.907
Media AREC/Corte=0.985	2175	2390	4618	2452	6440/0.984	3015
Intervalo de Confianza AREC	2017-2334	2226-2555	4471-4765	2270-2634	6337-6543	2900-3130
Cantidad Promedio de Individuos Simulados	32525	35354	68410	34838	90000	42605
Entropía Final Promedio	6.387	6.474	7.143	6.755	7.520	6.879

Como se ve, los criterios de corte más restrictivos produjeron, ahora sí, comportamientos de los estimadores de riqueza ARE y AREC similares a los obtenidos para el otro conjunto de muestras Para la muestra Err23 no alcanzó el "loop" de individuos simulados para llegar a los valores de corte propuestos. Los umbrales elegidos en este caso, 0.015 para la proporción de singletones y 0.985 para la cobertura, revelan que el valor de corte para obtener estimaciones de parecida eficiencia, sobre distintos conjuntos muestrales, puede variar. Como esta variación puede, a su vez, resultar en un cambio significativo en la cantidad de individuos simulados en cada caso, no resulta conveniente fijar un número determinado de ellos como criterio para terminar la simulación. Al respecto parece más adecuado estudiar la sensibilidad de la estimación utilizando distintos umbrales para la proporción de singletones y la cobertura como aquí se ha hecho.

5.6 USO DEL COEFICIENTE DE VARIACIÓN (ARECV)

- *La variabilidad de las frecuencias*

La llamada "frecuencia de la frecuencia" es usada en la simulación cuando, en cada paso, se actualiza la distribución en OTUs. Que una OTU tenga un solo elemento significa que su frecuencia de aparición en la muestra de individuos es 1. La cantidad de OTUs que tienen esa frecuencia es la frecuencia con que se presenta, precisamente, la frecuencia 1. Por supuesto, de las cantidades absolutas puede pasarse a las relativas, que son las efectivamente usadas para simular la condición de cada nuevo individuo. Como demuestra Good en [21], la distribución de la "frecuencias de la frecuencias" es la que permite evaluar la probabilidad de un individuo de pertenecer a una especie nueva a través del cociente entre el número de especies que se presentan una sola vez sobre el total de individuos de la muestra. En cada paso de la simulación, esta chance se estima por medio de $\hat{T}_i = \frac{n°sgletones}{i-1}$ citada en (5.1.1), donde cada singletón en la muestra de i-1 individuos está formado por un solo individuo que representa a su especie. Cuando esta estimación de probabilidad de especie nueva se modifica al multiplicarla por la cobertura para corregirla, se obtiene

$p_{ns} = \hat{T}_{i+1}\hat{C} = \frac{f_1}{i}(1 - \frac{f_1}{i})$ Se busca así modelar el hecho de que la muestra, dado su escaso tamaño en relación con la población, "cubre" solo una parte de la distribución de especies real.

Al realizar la simulación incorporando individuos de a uno a la muestra, e incrementando paulatinamente el número de especies en el medio simulado, se ha tenido en cuenta hasta aquí la cobertura. Sin embargo, no se ha apreciado específicamente la variabilidad de las "frecuencias de las frecuencias" en cada estado del proceso aleatorio. En cada iteración las frecuencias correspondientes a singletones, dubletones, etc, sufren una variación respecto de las anteriores o bien porque se incorpora una nueva especie o bien porque se agrega un nuevo individuo a una especie ya antes presente. Como consecuencia, la variabilidad del conjunto de "frecuencias de frecuencias" cambia de iteración en iteración pudiendo ser en el paso actual mayor o menor que en el anterior. En la medida en que se incorporan más individuos a la muestra lo que deberá ocurrir con el modelo de simulación para ajustar a la realidad es que la variabilidad muestral de las "frecuencias de las frecuencias" vaya convergiendo a su variabilidad poblacional.

La población contiene en la realidad una parte de individuos de cada especie. Si N es el tamaño desconocido poblacional y N_j la cantidad desconocida de individuos de la especie j entre esos N, se tiene que la proporción $p_j = \frac{N_j}{N}$ representa la probabilidad de que un individuo tomado al azar pertenezca a la especie j. Estas probabilidades, consideradas para todo j, presentan una distribución que varía según las características del medio considerado pues depende de la relación uniformidad-dominancia entre las especies. En un extremo, todas las especies podrían tener igual probabilidad de aparecer al seleccionar al azar un individuo, como también podría haber especies frecuentes y otras raras. El desvío estándar de estas frecuencias poblacionales (no de la variable de cuyos casos son probabilidad) se calcula como

$$\sigma = \left(\frac{\sum_{j=1}^{S} (p_j - \bar{p})^2}{S} \right)^{\frac{1}{2}} \qquad (5.6.1)$$

con

$$\bar{p} = \frac{\sum_{j=1}^{S} \frac{N_j}{N}}{S} = \frac{1}{S} \qquad (5.6.2)$$

y S cantidad de especies en la población. Claramente, si la distribución de las especies fuese exactamente uniforme, la variabilidad de sus probabilidades sería medida por $\sigma = 0$. En [20] la variabilidad del conjunto de probabilidades es estimada por Chao y Lee, a partir de la muestra, para elaborar el estimador ACE según la fórmula dada en (3.2.10)

Las proporciones $p_j = \frac{N_j}{N}$ y también los números N y S, con las que debieran medirse, en realidad no se conocen y se pretende estimarlas a través de una muestra. En particular se desea estimar S, número de especies.

Se puede considerar entonces la siguiente alternativa. Cuando se simula incorporando individuos, la variabilidad del conjunto de proporciones de singletones, dubletones, etc, en el total de la muestra, tendrá que irse estabilizando con el incremento de la cobertura, al asociarse adecuadamente con la real distribución de la población. Tal estabilización puede modelarse por medio de un coeficiente que tome en cuenta las diferencias que muestra a muestra se presentan en la variabilidad. Para medir ésta, con carácter relativo a la muestra, se puede utilizar, en cada iteración, el coeficiente de variación de las proporciones de especies $k_i = \frac{\sigma_i}{\bar{p}_i}$ con el desvío estándar de la frecuencia de las especies en la muestra dado por $\sigma_i = \left(\frac{\sum_{j=1}^{S_i} (\hat{p}_j - \bar{p}_i)^2}{S_i} \right)^{\frac{1}{2}}$

y el promedio de las frecuencias $\bar{p}_i = \frac{\sum_{j=1}^{S_i} \frac{i_j}{i}}{S_i} = \frac{1}{S_i}$ donde S_i es la cantidad de especies distintas en la muestra de tamaño i y i_j es la cantidad de individuos de la especie j presentes en la muestra.

Se propone entonces calcular, en cada paso simulado, el coeficiente $A_i = 1 + (\frac{k_{i-1}}{k_i})$ para aplicarlo como factor multiplicador de la cobertura C_i,

$C\,\mathrm{mod}_i = C_i A_i$ y recalcular a continuación $p_{ns} = \hat{T}_i C\,\mathrm{mod}_i$. Para resolver el problema que plantea la posibilidad de que el producto $C\,\mathrm{mod}_i = C_i A_i$ pueda crecer de tal forma que resulte p_{ns} mayor que 1, lo que no debe ocurrir pues es la estimación corregida de una probabilidad, y para evitar que el cociente $(\frac{k_{i-1}}{k_i})$ conduzca a un valor de A_i demasiado grande que haga lenta la convergencia, se puede calcular $A_i = 1 + \frac{1}{\alpha}(\frac{k_{i-1}}{k_i})$ eligiendo adecuadamente el valor $\alpha \succ 0$. El resto de los pasos de la simulación se deben realizar como en 5.3. Se la denomina ARECV y el programa respectivo desarrollado en lenguaje R se incorpora al CD anexo bajo el nombre Pruebas2.R

- *Resultados*

La simulación ARECV que utiliza la cobertura y el coeficiente de variación se probó sobre el conjunto SRX008158. Los resultados se exhiben en la Tabla 5.6.1

Tabla 5.6.1

Muestra	N° Individuos	ACE	ARE	AREC	Cobertura Final	ARECV	Número de Individuos Simulados	Entropía Inicial
S85	1641	1102	997	938	0.94698	2008	4000	5.938
S86	8361	9595	6650	5754	0.90765	10640	15000	7.759
S87	6926	5964	4271	3834	0.93347	8044	14000	7.049
S88	6146	3755	3920	3552	0.93204	7323	13000	6.852
S89	6226	2755	3067	2882	0.95292	6250	13000	6.767
S90	8444	3544	3537	3307	0.95184	6663	15000	6.669
S91	6103	3757	3065	2928	0.94607	6003	13000	6.687
S92	5885	4048	3417	3163	0.93743	6693	12000	6.878

Se utilizó un valor $\alpha = 0.8$ y se juzgó que los valores obtenidos fueron demasiado elevados. Para el conjunto ERX009564 se realizó igual prueba con el mismo valor del parámetro α. Los resultados son los de la Tabla 5.6.2

Tabla 5.6.2

Muestra	N° Individuos	ACE	ARE	AREC	Cobertura Final	ARECV	Número de Individuos Simulados	Entropía Inicial
Err19	5011	2239	1735	1639	0.96508	3610	10000	5.922
Err20	5582	2458	1866	1846	0.96309	3914	11000	5.979
Err21	7637	4531	3508	3229	0.95224	7048	15000	6.640
Err22	3299	2125	1881	1664	0.94765	3594	7000	6.260
Err23	10371	6237	4942	4545	0.94774	9627	20000	6.974
Err24	5860	2129	2346	2256	0.95711	4449	11000	6.391

Aquí también las estimaciones de riqueza ARECV, con cobertura y coeficiente de variación, fueron demasiado elevadas. Se consideró que el ajuste del parámetro α requería pruebas con poblaciones simuladas o reales no disponibles y que la fórmula de estimación, a pesar de contener una idea promisoria, podía requerir algunas modificaciones posteriores. Por tal motivo se decidió abandonar por el momento este camino de estimación.

5.7 USO DE LA ENTROPÍA (AREN Y ARECE)

- *Efectos Observados sobre la Entropía Muestral*

A fin de obtener una forma de estimación que hiciera uso de la entropía, que es una de las medidas más usadas para evaluar diversidad biológica, se pensó en modificar en cada iteración el valor de la estimación de la probabilidad de especie nueva, utilizando un coeficiente que la tuviera en cuenta. Al respecto; de las observaciones recogidas en las pruebas de simulación detalladas en 5.5, se extrajeron dos conclusiones importantes. La primera registra que la entropía crece según aumenta la cantidad de individuos simulados y la segunda constata que la diferencia entre los valores de dos pasos sucesivos de la simulación va disminuyendo conforme aumentan los casos simulados y se va convergiendo a los valores de riqueza poblacionales. Es decir, la entropía crece buscando converger a su valor final poblacional.

Los efectos comentados fueron entonces utilizados para proponer, en cada iteración, el coeficiente $B_i = 1 + H_i \left| H_i - H_{i-1} \right|$. De esta forma se corrige la

probabilidad de especie nueva según $\hat{T}_i corr = B_i \hat{T}_i$. Alternativamente la cobertura también puede modificarse de acuerdo a $C\,\mathrm{mod}_i = C_i B_i$ y la probabilidad de especie nueva corregirse con $p_{ns} = \hat{T}_i C\,\mathrm{mod}_i$. Se busca así tener en cuenta, en cada paso de simulación, el valor de la entropía que debe crecer con la cantidad de individuos simulados, junto con la disminución de las diferencias entre las entropías del paso actual y del anterior conforme avanza el número de pasos. Para calcular la entropía en cada paso de la simulación se utiliza la fórmula $H_i = -\sum_{j=1}^{S_i} \hat{p}_j \log \hat{p}_j$ con $\hat{p}_j = \frac{i_j}{i}$ en cada iteración. Finalmente con el objetivo de contar con una herramienta de eventual regulación de la influencia de la entropía en la modificación de la probabilidad de especie nueva se introduce un parámetro $\beta \geq 0$ en el cálculo del coeficiente. Así $B_i = 1 + \beta H_i \left| H_i - H_{i-1} \right|$ Con esta versión de B_i se calcula como fue indicado $\hat{T}_i corr$ en el caso de ARE, y $C\,\mathrm{mod}_i$ y p_{ns} en el caso de AREC. En ambos algoritmos la secuencia de programación continúa como fue señalado en 5.2 y 5.3 respectivamente. Las simulaciones, denominadas AREN y ARECE respectivamente, fueron programadas en lenguaje R e incorporadas al CD anexo con los nombres Prueba_Turing_Entropia.R y Prueba_Cobertura_Entropia.R

- *Resultados*

Las pruebas efectuadas sobre el conjunto de muestras SRX008158 con el parámetro $\beta = 0.85$ condujeron a los valores que se exhiben en la Tabla 5.7.1

Tabla 5.7.1

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
N° Individuos	1641	8361	6926	6146	6226	8444	6103	5885
Entropía Inicial	5.938	7.759	7.049	6.852	6.767	6.669	6.687	6.878
Sobs	541	3575	2273	2030	1715	2040	1659	1842
CHAO	834	6351	4080	3363	2502	3180	2691	3110
ACE	1102	9525	5964	3755	2755	3544	3757	4048
Número de Simulaciones	7	7	7	7	7	7	7	7
Confianza del Intervalo	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Media ARE Tabla 5.5.5	1288	9851	6015	5567	3759	4384	3924	4800
Media AREN/Corte=0.03	1298	9826	6164	5539	3865	4409	3953	4774
Intervalo de Confianza AREN	1188-1408	9536-10116	5859-6469	5200-5879	3714-4016	4192-4626	3833-4074	4525-5023
Cantidad Promedio de Individuos Simulados	10789	60000	53085	48129	31390	36194	33228	40911
Entropía Final Promedio	6.294	8.321	7.551	7.381	7.122	7.018	7.087	7.343
Media AREC Tabla 5.5.5	1088	8165	5355	4743	3482	4223	3586	4286
Media ARECE/ Corte=0.975	1123	8168	5338	4612	3471	4105	3572	4242
Intervalo de Confianza ARECE	1047-1199	8065-8271	5222-5454	4407-4817	3335-3607	3980-4230	3377-3766	4105-4378
Cantidad de Individuos Simulados	10130	60000	50910	42988	30857	37607	33425	40537
Entropía Final Promedio	6.188	8.132	7.428	7.221	7.033	6.948	6.993	7.230

Aquí la estimación AREN resultó, para todas las muestras, mayor que la aportada por ACE. Sin embargo en la comparación con los valores hallados al usar ARE, la corrección propuesta no alcanzó cambios relevantes. La estimación ARE recayó siempre en el intervalo de confianza hallado para la estimación AREN. El

parecido se reitera entre las estimaciones AREC y ARECE, realizadas con cobertura con y sin entropía respectivamente.

A fin de intentar una mejora en las estimaciones se decidió elevar el valor del parámetro β. Para el conjunto ERX009564 con $\beta = 1.0$ se obtuvieron los siguientes resultados expresados en la Tabla 5.7.2

Tabla 5.7.2

Muestra	Err19	Err20	Err21	Err22	Err23	Err24
N° Individuos	5011	5582	7637	3299	10371	5840
Entropía Inicial	6.132	6.212	6.825	6.431	7.200	6.613
Sobs	957	1066	1930	997	2664	1358
CHAO	1821	1786	3374	1687	4937	2104
ACE	2239	2458	4531	2125	6237	2129
Número de Simulaciones	7	7	7	7	7	7
Confianza del Intervalo	0.95	0.95	0.95	0.95	0.95	0.95
Media ARE Tabla 5.5.6	1856	2102	4357	2338	6244	2732
Media AREN/Corte=0.03	1917	2149	4289	2475	6235	2687
Intervalo de Confianza AREN	1831-2003	2032-2266	4132-4448	2377-2572	6065-6406	2536-2838
Cantidad Promedio de Individuos Simulados	15449	16968	35328	20853	52591	19881
Entropía Final Promedio	6.365	6.471	7.162	6.823	7.564	6.865
Media AREC Tabla 5.5.6	1810	2035	4076	2167	5668	2579
Media ARECE/Corte=0.975	1851	2056	4100	2145	5767	2639
Intervalo de Confianza ARECE	1780-1922	1941-2171	3951-4249	2008-2281	5681-5853	2521-2757
Cantidad Promedio de Individuos Simulados	16574	18571	38146	19515	54767	23533
Entropía Final Promedio	6.351	6.436	7.120	6.701	7.493	6.832

Se observa que la modificación por entropía propuesta AREN no produce cambios importantes respecto de los resultados de la simulación ARE permaneciendo, para un valor de corte de 0.03, también por debajo de las estimaciones ACE. Para la estimación con cobertura y entropía ARECE, con corte a 0.975, en general se mantiene la relación observada anteriormente con las estimaciones ARE y ACE. Por otra parte las estimaciones ARE y AREC ofrecen la ventaja de no depender de la elección, un tanto arbitraria, del parámetro β. De todas formas y a efecto de confirmar el comportamiento

similar de la estimación con entropía, aún cuando iguales parámetros de corte hayan permitido elevar los valores estimados por sobre ACE, se realizó una prueba con el conjunto SRX008158 para el mismo valor $\beta = 1.0$ Los resultados se muestran en la Tabla 5.7.3

Tabla 5.7.3

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
N° Individuos	1641	8361	6926	6146	6226	8444	6103	5885
Entropía Inicial	5.938	7.759	7.049	6.852	6.767	6.669	6.687	6.878
Sobs	541	3575	2273	2030	1715	2040	1659	1842
CHAO	834	6351	4080	3363	2502	3180	2691	3110
ACE	1102	9525	5964	3755	2755	3544	3757	4048
Número de Simulaciones	7	7	7	7	7	7	7	7
Confianza del Intervalo	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Media ARE Tabla 5.5.5	1288	9851	6015	5567	3759	4384	3924	4800
Media AREN/Corte=0.03	1294	9880 /0.043	5997	5393	3735	4401	3860	4781
Intervalo de Confianza AREN	1178-1410	9711-10049	5797-6198	5172-5615	3681-3790	4310-4492	3740-3980	4609-4952
Cantidad Promedio de Individuos Simulados	10500	60000	51269	45874	29856	36381	32127	41074
Entropía Final Promedio	6.296	8.329	7.538	7.364	7.104	7.006	7.054	7.341
Media AREC Tabla 5.5.5	1088	8165	5355	4743	3482	4223	3586	4286
Media ARECE/ Corte=0.975	1069	8102	5266	4634	3456	4189	3623	4254
Intervalo de Confianza ARECE	1000-1138	8032-8173	5116-5416	4540-4728	3371-3540	4125-4253	3521-3725	4072-4436
Cantidad de Individuos Simulados	9036	60000	50036	43034	30528	38864	34314	40519
Entropía Final Promedio	6.181	8.125	7.417	7.229	7.033	6.973	7.006	7.226

Las estimaciones con la modificación propuesta por entropía oscilan, muestra a muestra, alrededor de las obtenidas por la simulación ARE sin corrección resultando además, en todos los casos, mayores que las realizadas por el índice ACE. Esto confirma el comportamiento preanunciado por la prueba sobre el otro conjunto

muestral y permite concluir que la simulación AREN ofrece también una alternativa para mejorar la estimación de la riqueza poblacional. A su vez sugiere una dirección de continuidad del trabajo, incorporando adicionalmente una corrección del sesgo negativo para la estimación de la entropía [17] y buscando el ajuste del parámetro β cuyas características debieran ser estudiadas con datos de poblaciones simuladas y reales.

6- CONCLUSIONES

6.1 CONSIDERACIONES GENERALES

El trabajo realizado ha permitido evaluar las serias dificultades que presenta la estimación de la diversidad microbiológica entendida como riqueza y forma de distribución de las especies o, más generalmente, de los taxones. Aún adoptando una metodología en particular, tal como lo es el análisis del gen "marcador" 16S rRNA, existe un amplio panorama de enfoques posibles que sugieren desde ajustes al proceso inicial de las secuencias que utilicen distintos patrones de alineado, diferentes criterios de filtrado para eliminar "gaps" o modelos alternativos para establecer la distancia genética; hasta variadas formas de estimación estadística tales como las no paramétricas o las realizadas por rarefacción. Toda esta batería de técnicas disponibles no logra, por lo general, resolver el problema de la subestimación de la riqueza real de una comunidad.

A favor de las crecientes rapidez de cálculo y capacidad de almacenamiento de datos que ofrece la tecnología computacional, las técnicas aquí propuestas logran resultados eficientes que mejoran la estimación en el contexto paradojal en que se cuenta con gran cantidad de datos y, sin embargo, estos no son suficientes para resolver, utilizando los procedimientos estadísticos usuales, la incertidumbre sobre los comportamientos poblacionales. Así la simulación se revela en el caso como una herramienta que permite explotar la abundancia de datos con que se cuenta y colaborar en la resolución de tal incertidumbre. En la medida en que los procedimientos sugeridos convergen a valores estimados de riqueza poblacional, van poniendo de manifiesto una distribución de taxones que estaba implícita en las muestras, cuya cantidad real de información es establecida por la entropía "final". Al menos este aspecto del conocimiento de la población queda entonces al descubierto junto con una estimación más "realista" de la riqueza. Este enfoque de explotación de los datos y, a partir de ellos, de descubrimiento de conocimiento es precisamente una de las ventajas que ofrece la minería de datos por sobre las estimaciones paramétricas, y aún sobre las no paramétricas, que no pueden captar la totalidad de la información presente. Los procedimientos de estimación construidos logran mayor grado de revelación de la información contenida implícitamente en los datos, al hacerla explícita a través de los patrones de riqueza y distribución hallados.

La estimación por simulación muestra la potencia y profundidad de la idea concebida por Alan Turing al producir un valor que, bajo la condición de elección de parámetros de corte adecuados, mejora los resultados de las inferencias realizadas por los distintos métodos desarrollados para resolver la cuestión. En las Figura 6.1.1.a y 6.1.1.b puede apreciarse la mejoría resultante para los conjunto muestrales utilizados en este trabajo.

Figura 6.1.1.a

Desempeño de Estimadores para el Conjunto SRX008158

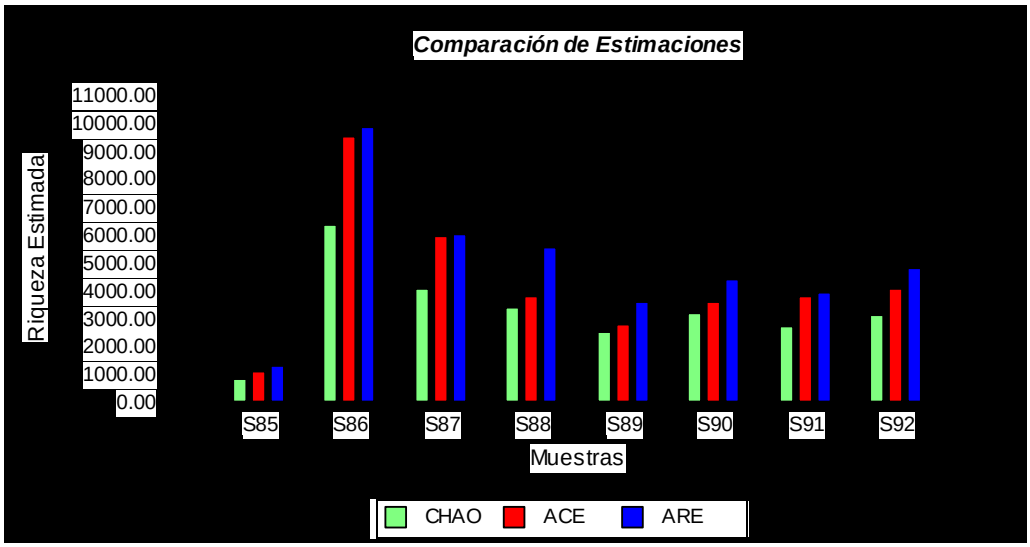
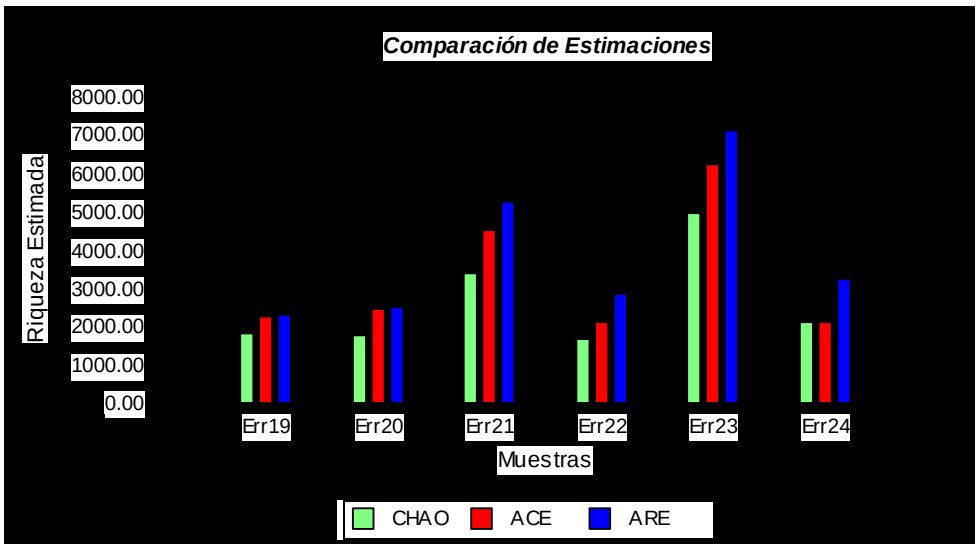


Figura 6.1.1.b

Desempeño de Estimadores para el Conjunto ERR009564



Complementariamente la relación porcentual de crecimiento que se ilustra a su vez en las Figuras 6.1.2.a y 6.1.2.b permite concluir que la estimación por el Algoritmo de Recuento de Especies ARE, ofrece la perspectiva de mejorar en todos los casos el desempeño de las técnicas no paramétricas.

Figura 6.1.2.a

Mejora Porcentual de la Estimación ARE para el Conjunto SRX008158

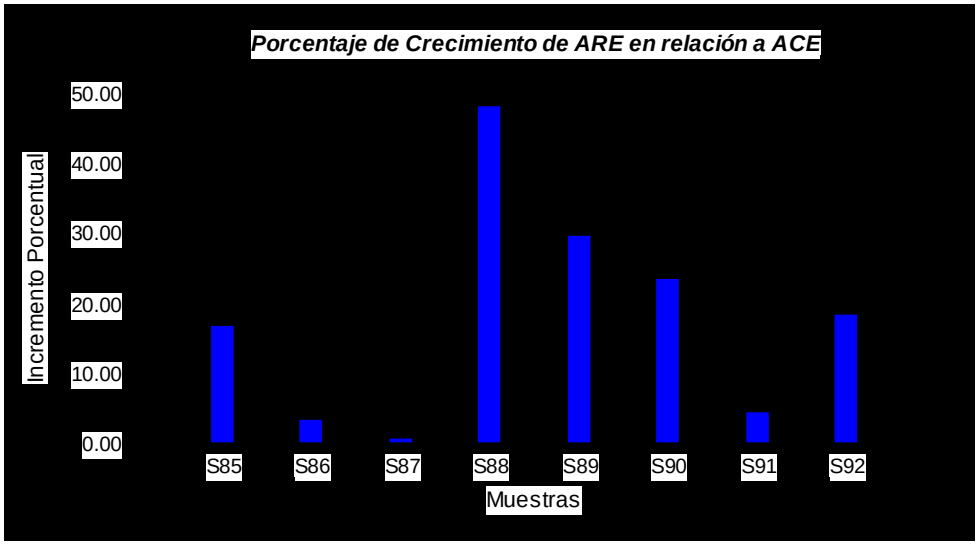
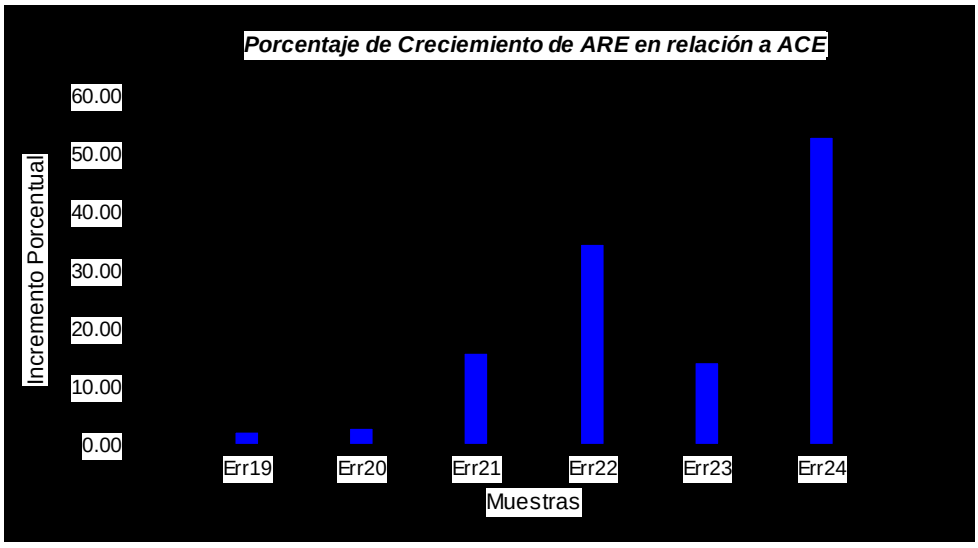


Figura 6.1.2.b

Mejora Porcentual de la Estimación ARE para el Conjunto ERR009564



En cuanto al procedimiento de simulación AREC que utiliza el concepto de cobertura, su importancia radica en que refuerza la idea de Turing atendiendo a la parte de la distribución de taxones que puede efectivamente estimarse conforme se

incrementa el número de casos simulados. Con una adecuada elección para la cobertura de corte plantea una eficiencia similar a la de ARE en la estimación de riqueza aunque la convergencia es más lenta, lo que se traduce en tiempos mayores de ejecución de programa. Esto se hace evidente cuando se comparan los resultados para similares esfuerzos de cómputo -o equivalentemente similares esfuerzos de muestreo simulado- según se observa en las Figuras 6.1.3.a y 6.1.3.b que corresponden al análisis de los conjuntos de muestras seleccionados para pruebas.

Figura 6.1.3.a

Estimaciones ARE y AREC para el Conjunto SRX008158

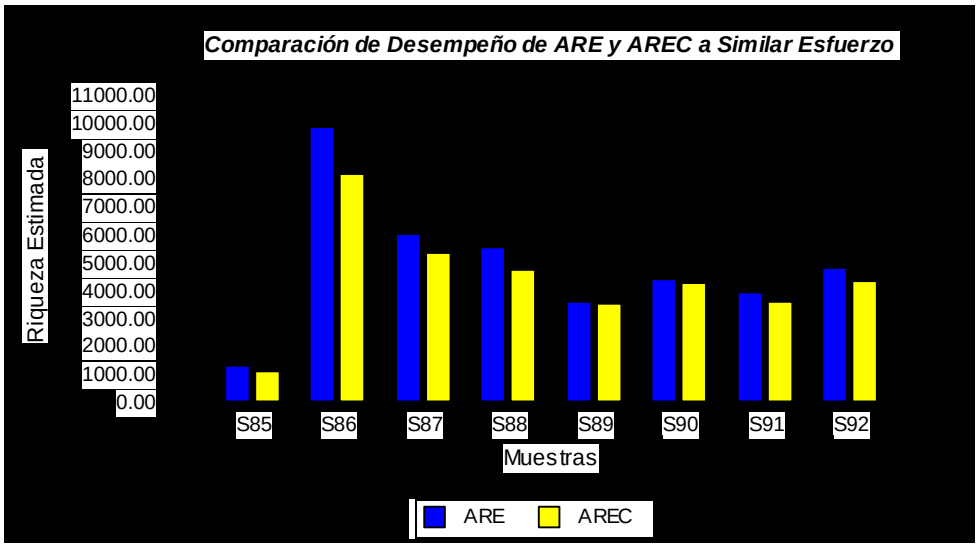
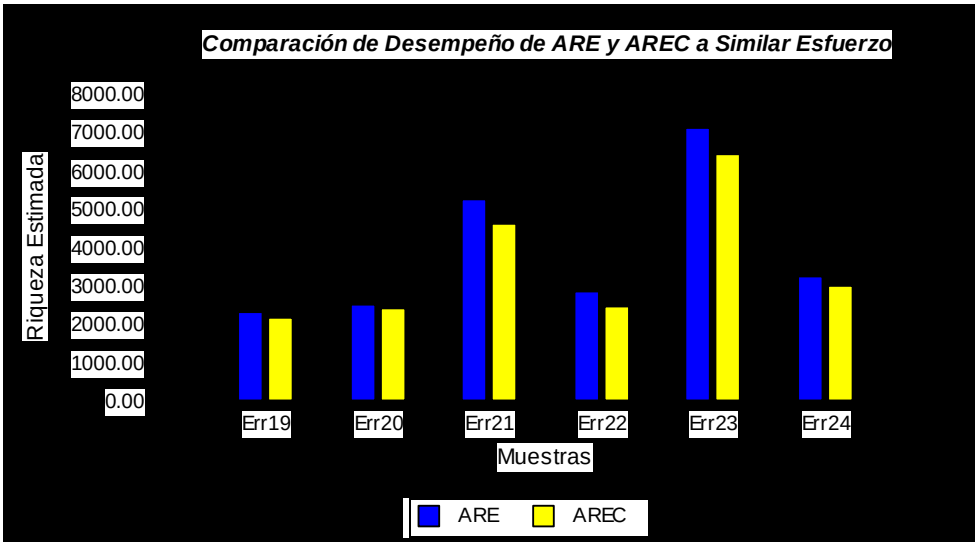


Figura 6.1.3.b

Estimaciones ARE y AREC para el Conjunto ERR009564



Los valores inferiores obtenidos mediante la simulación con cobertura aumentarán a condición de que se incremente el número de casos simulados al establecer un criterio de corte más exigente. Por otra parte el suavizado ARECS1 descripto en 5.4, que en las pruebas preliminares produjo estimaciones levemente superiores a las de la simulación con cobertura, ofrece una alternativa a explorar a fin de disminuir el esfuerzo de muestreo simulado y no obstante lograr valores más ajustados.

Si bien hasta aquí ninguno de los algoritmos planteados cuenta con una prueba matemática de convergencia o una comprobación empírica sobre toda una población real o incluso simulada, su construcción se sustenta en pruebas de índole matemática sobre el carácter de estimador que poseen las fórmulas de Turing y de Cobertura empleadas. Ver [21] y [20] respectivamente. Además la aplicación sobre los conjuntos muestrales seleccionados permite confiar razonablemente en su desempeño eficiente en similares condiciones. No es ése el caso del método de simulación ARECV desarrollado en 5.6, que incorpora el coeficiente de variación de las frecuencias en la corrección de la probabilidad de especie nueva. La fórmula empleada en tal variante es una construcción basada en una idea que puede ser promisoria pero que requiere mayores estudios para ser aplicada. Los resultados preliminares obtenidos sugieren explorar, al menos, modificaciones en la estructura de la fórmula y nuevas pruebas sobre muestras.

Durante el desarrollo del trabajo se procuró aclarar la significación y los alcances del concepto de entropía en relación con la diversidad biológica. Por ese motivo se prefirió ese índice de diversidad a otros que como el de Simpson también estaban disponibles. En primer término las consideraciones efectuadas en 4.2 permiten establecer que dado un cierto valor de entropía muestral, la existencia de una diferencia entre la cantidad observada de taxones y la que correspondería a la entropía calculada, revela una probable pérdida de uniformidad, una presencia de dominancia de unas especies por sobre otras en la población. Es claro que tal cuestión debe tratarse con cuidado pues no puede evaluarse “a priori” el grado de probabilidad que haya tenido la particular selección muestral. Sin embargo al incrementarse el tamaño muestral, cubriendo una mayor proporción del espacio de probabilidad, por vía de la simulación planteada, la entropía calculada va representando más adecuadamente a la poblacional. En tal caso va perdiendo sentido comparar el valor actual de la entropía y la cantidad de especies que pudiera corresponder a una distribución uniforme, pues se cuenta ahora

con una forma más aproximada y probable de la distribución poblacional real de las especies.

Con la intención de reunir en una sola forma de estimación la simulación realizada utilizando la fórmula de Turing, con y sin corrección por cobertura, y la cantidad de información de las sucesivas muestras calculada por la entropía, se construyeron los procedimientos AREN y ARECE de estimación con entropía que se presentó en 5.7 Los resultados alcanzados muestran que para elecciones adecuadas del parámetro de ajuste se presentan solo leves variaciones respecto de los obtenidos sin corregir por entropía. Para la estimación AREN las Figuras 6.1.4.a y 6.1.4.b evidencian tal conclusión, observándose tanto incrementos como decrementos porcentuales.

Figura 6.1.4.a

Efecto de la Corrección por Entropía sobre ARE para el Conjunto SRX008158

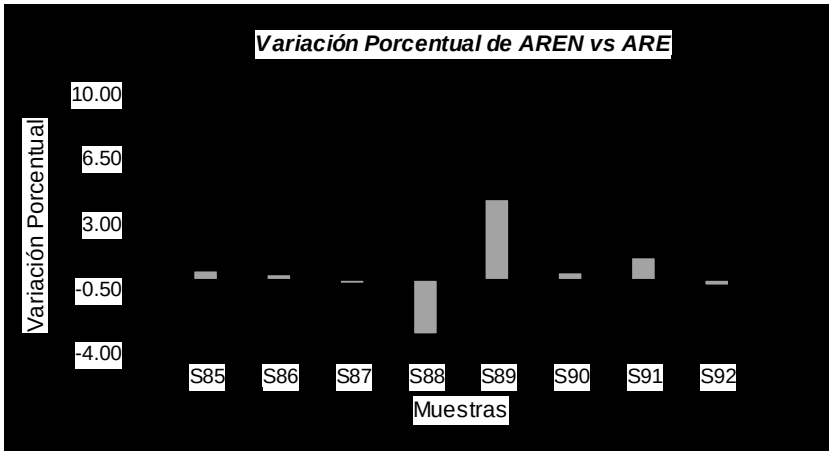
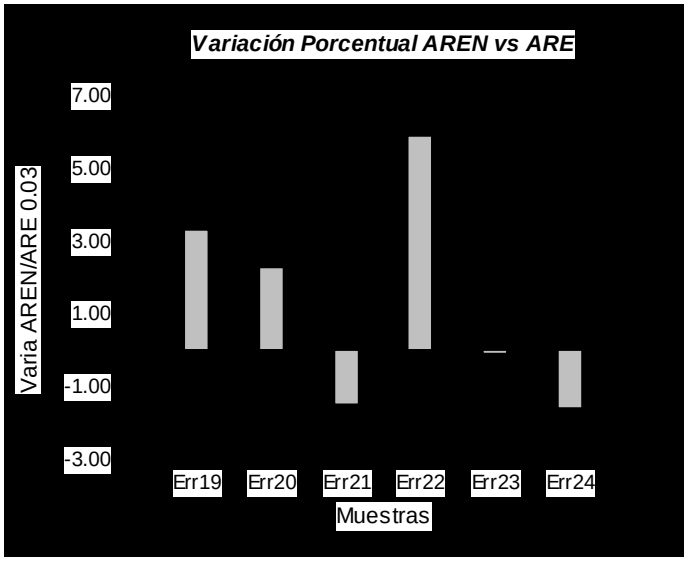


Figura 6.1.4.b

Efecto de la Corrección por Entropía sobre ARE para el Conjunto ERR009564



No obstante lo expuesto queda por explorar el efecto que tendría sobre la estimación propuesta la corrección por sesgo de los valores estimados de entropía en cada iteración.

Las variaciones registradas cuando se aplicó la simulación con cobertura y corrección por entropía ARECE respecto de la versión sin corrección AREC, fueron también pequeñas pero mostraron una tendencia más constante en su signo, lo que se ve en la Figuras 6.1.5.a y 6.1.5.b

Figura 6.1.5.a

Efecto de la Corrección por Entropía sobre AREC para el Conjunto SRX008158

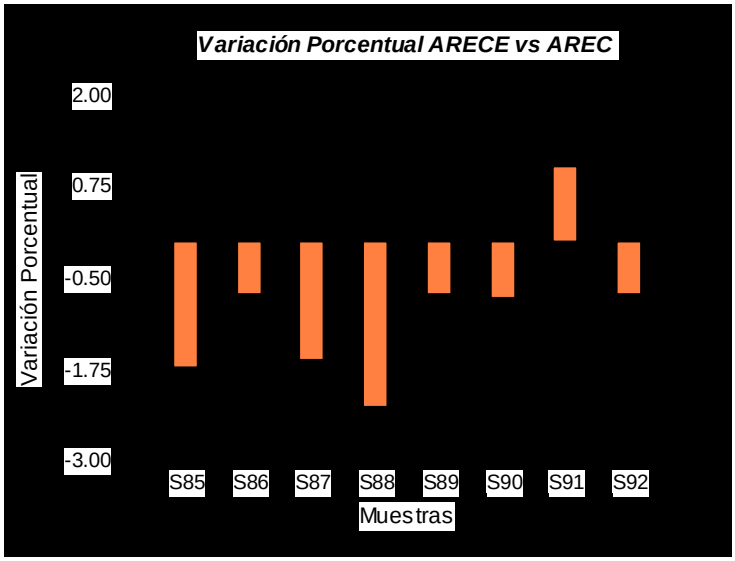
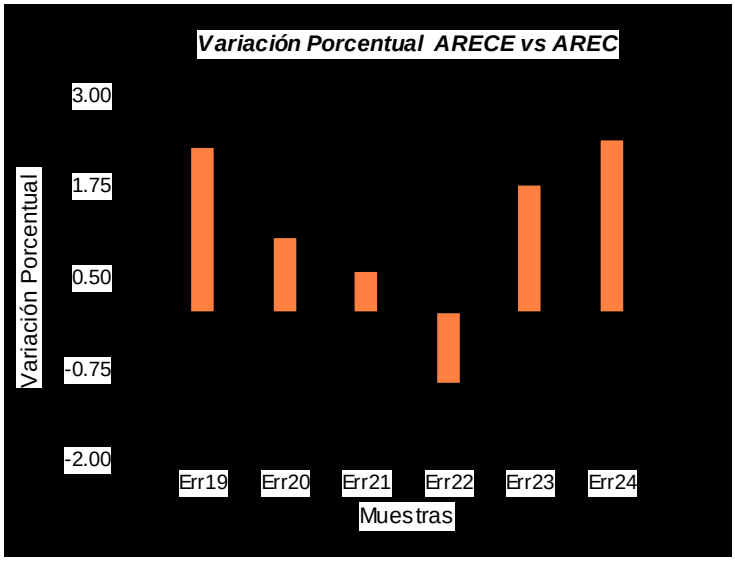


Figura 6.1.5.b

Efecto de la Corrección por Entropía sobre AREC para el Conjunto ERR009564



La mayor constancia en el signo de la variación y la diferencia en ese signo cuando se analizan los dos conjuntos muestrales seleccionados, sugieren la necesidad de más pruebas sobre otros conjuntos muestrales, que además incorporen la corrección por sesgo de la estimación de entropía y utilicen un valor de cobertura de corte más exigente combinado con un ajuste más sensible del parámetro β citado en 5.7. Esas experiencias debieran servir para determinar un intervalo de variación de tal parámetro y podrían complementarse con pruebas del desempeño del algoritmo con cobertura y corrección por entropía sobre poblaciones simuladas.

6.2 PERSPECTIVAS

Al considerar los aspectos estadísticos y computacionales a partir del trabajo desarrollado en el contexto del análisis de genes marcadores, se abren tres líneas de investigación para consolidar la mejora obtenida en las estimaciones de riqueza de poblaciones microorganicas. La primera de ellas se relaciona con posibles modificaciones en las construcciones de las fórmulas de estimación utilizadas, que incorporen otras cantidades útiles avaladas por propiedades matemáticamente demostradas y ensayadas sobre nuevos conjuntos muestrales. La segunda línea de investigación se refiere a los ensayos de los estimadores aquí construidos sobre poblaciones simuladas y, en la medida que vayan acumulándose datos reales de una misma población, también a pruebas sobre los mismos. Esto involucra además la forma de acelerar los tiempos de cálculo y de utilizar el cómputo en paralelo. Una tercera línea de investigación, que quizás permita generar también nuevas ideas, la constituye la construcción de las pruebas matemáticas formales de la convergencia de los algoritmos desarrollados, que deberán establecer alcances y restricciones para los mismos.

Más allá de las tres posibilidades arriba citadas, es claro que la elección de otras alternativas en los procesos previos del ADN, tales como alineamiento, filtrado o construcción de la matriz de distancias, puede eventualmente producir mejoras en la estimación de la riqueza real de la población. Finalmente, también se podría intentar la evaluación de la riqueza poblacional al analizar las secuencias obtenidas por aplicación de la técnica WGS citada en 2.2, lo que requeriría nuevas construcciones de estimadores y algoritmos distintos de los aquí presentados.

BIBLOGRAFIA

[1] [2008] Youssef, N y Elshahed, M. “Species richness in soil bacterial communities: A proposed approach to overcome sample size bias”. Journal of Microbiological Methods. 75 86-91.

[2] [2001] Hughes, J, Hellmann, J, Ricketts, T y Bohannan, B. “Counting the uncountable: statistical approaches to estimating microbial diversity”. Appied and Environmental Microbiology. 4399-4406.

[3] [1998] Durbin, R, Eddy, S, Krogh, A y Mitchison, G. Biological Sequence Analysis. Cambridge University Press.

[4] [2007] Gross, L. “Untapped Bounty: Sampling the Seas to Survey Microbial Biodiversity”. PLoS Biology/ Volume 5/Issue 3/e85

[5] [2009] Guazzaroni, M.E, Beloqui, A, Golyshin, P y Ferrer, M. “Metagenomics as a new technological tool to gain scientific knowledge”. World Journal Microbiologic Biotechnology. 25:945-954

[6] [2006] Schloss, P. y Handelsman,J. “Toward a census of bacteria in soil”. PLoS Computational Biology. Volume 2.

[7] [2009] White, J, Nagarajan, N, Pop, M. “Statistical methods for detecting differentially abundant features in clinical metagenomic samples”. PLoS Computational Biology. Volume 5.

[8] [2009] Brady, A y Salzberg, S. “Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models”. Nature Methods. Vol. 6 N° 9.

[9] [1996] D.M. Hillis, C. Moritz, B.K. Mable. Molecular Systematics. Second Edition, Sinauer Associates, Inc. Publishers. Sunderland, MA. USA

[10] [2005] Schloss, P y Handelsman, J. “ Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness”. Applies and Environmental Microbiology. Pgs. 1501-1506

[11] [2001] Everitt, B, Landau, S, Leese, M. Cluster Analysis. Fourth Edition. Arnold.

[12] [2005] O’Hara, R. “Species richness estimators: how many species can dance on de head of a pin. Journal of Animal Ecology. 74, 375.386

[13] [1949]. Shannon, C. The mathematical theory of communication. Illini Books edition.

[14] [2003] Hill, T, Walsh, K, Harris, J yMoffett, B. “Using Ecological Diversity Measures with Bacterials Communities”. FEMS.Microbiology Ecology 43 1-11

[15] [2004] Magurran, A. Measuring Biological Diversity. Blackwell Science Ltd

[16] [1993] Bunge, J. y Fitzpatrick, M. “Estimating the Number of Species: A Review”. Journal of American Statistical Association. Vol. 88. N° 421. pp. 364-373.

[17] [2003] Chao, A y Shen, T. “Nonparametric estimation of Shannon’s index of diversity when there are unseen species in sample”. Environmental and Ecological Statistics 10, 429-443.

[18] [2001] Hughes, J, Hellmann, J, Ricketts, T y Bohannan, B. “Counting the uncountable: statistical approaches to estimating microbial diversity”. Appied and Environmental Microbiology. 4399-4406.

[19] [1984] Chao, A. “Nonparametric estimation of the number of classes in a population”. Scand J Statist 11: 265-270.

[20] [1992] Chao, A y Lee, S. “Estimating the Number of Classes via Sample Coverage”. Journal of American Statistical Association. Volume 87. Issue 417.

[21] [1953] Good, I. “The Population Frequencies of Species and Estimation of Population Parameters”. Biometrika. Vol 40 N° 3/4.

[22] [2005] Hughes,J y Hellman, J. “The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity”. Methods in Enzymology. Vol 397.

[23] [2001] Gotelli, N y Colwell,R. “Quantifying biodiversity: procedures and pitfalls in measurement and comparison of species richness” Ecology Letters. 4: 379-391

[24] [1978] Efron, B. “Computers and theory of statistics: thinking the unthinkable”. Technical Report N° 39. Division of Biostatistics. Stanford University

[25] [2009] Tellinghuisen, J. “The Least_Squares Analysis of Data from Binding and Enzyme Kinetics Studies: Weights, Bias, and Confidence intervals in Usual and Unusual Situations”. Methods in Enzymology. Volume 467. Pgs.500-527.

[26] [1982] Currie, D. “Estimating Michaelis-Menten Parameters: Bias, Variance and Experimental Design”. Biometrics. Vol. 38, N° 4 Pgs. 907-919.

[27] [1996] Swofford, D. Olsen, G. Waddell, P. y Hillis, D. Molecular Systematics. Chapter 11. Phylogenetic Inference. Second edition. Edited by David M. Hillis, Craig Moritz, and Barbara K. Mable.

[28] [2005] Schloss, P y Handelsman, J. “ Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness”. Applies and Environmental Microbiology. Pgs. 1501-1506.

[29] [2010] Schloss, P. “The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-based Studies”. PLoS Computational Biology 6(7): e1000844. doi:10.1371/Journal.pcbi.1000844.

[30] [2007] Roesch, L, Fulthorpe, R, Riva, A, Casella, G, Hadwin, A, Kent, A, Daroub, S, Camargo, F, Farmerie, W y Triplett, E. "Pyrosequencing enumerates and contrasts soil microbial diversity". The ISME Journal. 1, 283-290.

[31] [2010] Hollister, E, Engledow, A, Hammett, A, Provin, T, Wilkinson, H. y Gentry, T. "Shifts in microbial community structure along an ecological gradient of hypersaline soils and sediments". The ISME Journal. 1-10.

[32] http://www.ncbi.nlm.nih.gov/

[33] [2009] Schloss, P.D., et al., Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75(23):7537-41

[34] http://evolution.genetics.washington.edu/phylip.html

[35] [2008] InfoStat- versión 2008. Manual del Usuario. Grupo InfoStat, FCA, Universidad Nacional de Córdoba. Primera Edición, Editorial Brujas Argentina.

[36] http://www.r-project.org/

[37] [2002] Chao, A. y Bunge, J. "Estimating the Number of Species in a Stochastic Abundance Model". Biometrics 58 Pgs. 531-539

[38] [2010] Parks, D y Beiko, R. "Identifying biologically relevant differences between metagenomic communities". Bioinformatics Advance Access published February 3. The Oxford University Press.

[39] [2007] Raes, J, Foerstner, K. U y Bork, P. "Get the most out your metagenome: computacional analysis of environmental sequence data". Current Opinion in Microbiology. 10:490-498

[40] [2009] Klimke, W, Agarwala, R, Badretdin, A, Chetvernin, S, Ciufo, S, Fedorov, B, Kiryutin, B, O'Neill, K, Resch, W, Resenchuk, S, Schafer, S, Tolstoy, I y Tatusova, T. "The national center for biotechnology information's proteins clusters database". Nucleic Acids Research. Vol. 37.

[41] [2003] Chao, A. "Species Richness Estimation". Technical Report. Institute of Statistics. National Tsing Hua University.

[42] [2004] Colwell, R, Mao, Ch y Chang, J. "Interpolating, extrapolating, and comparing incidence-based species accumulation curves". Ecology. 85(10) págs 2717-2727.

[43] [1945] Cramer, H. Mathematical Methods of Statistics. Ed. Almquivst & Wiksells.
[44] [1996] Gotelli, N y Graves, G Null Models in Ecology. Smithsonian Institution Press.

[45] [1997] Setubal, J y Meidanis, J. Introduction to Computational Molecular Biology. PWS Publishing Company.

[46] [2006] Grasso, D. “Metagenómica: un viaje a las estrellas”. Revista Argentina de Microbiología. Volumen 38. N° 4.

[47] [2005] Ewens, W y Grant,G. Statistical Methods in Bioinformatics: An Introduction. Springer Science+Business Media, Inc.

[48] [2006] Hong, S, Bunge, J, Jeon, S, Epstein, S. “Predicting microbial species richness”. PNAS. Vol. 103 N°1 Págs. 117-122

[50] [1988] Canavos, G. Probabilidad y Estadística. McGraw-Hill

[51]- [2006] Borodovsky, Mark y Ekisheva, Svetlana. Problems and Solutions in Biological Sequence Analysis. Cambridge University Press. .

[52] [2005] Mao, Ch, Colwell, R y Chang, J. “Estimating the species accumulation

[53] [2006] Rodríguez Brito, B, Rohwer, F. y Edwards, R. “An application of statistics to comparative metagenomics”. BMC Bioinformatics. 7:162. 1-11.

[54] [2005] Salicrú, M, Vives, S, Ocaña,J. “Testing the homogeneity of diversity measures. A general framework”. Journal of Statistical Planning and Inference 132. 117-129.

[55] [2006] Smith, C y Pontius, J. “Jackniffe estimator of species with S-PLUS”. Journal of Statistical Software. Volume 15.

[56] [1984] Smith, E y van Belle, G. “Nonparametric estimation of species richness”. Biometrics 40. 119-129.

ANEXO

CONTENIDO DEL CD ADJUNTO

A- PROGRAMAS

1- SCRIPT_2E.R

El programa contiene las funciones necesarias para correr las simulaciones ARE, AREC, ARECS1, ARECS2 y ARECV. También contiene una función para calcular la entropía de una muestra. Se debe correr en la consola R antes de realizar las pruebas respectivas.

2- SCRIPT_TURING_ENTROPÍA.R

El programa contiene las funciones necesarias para correr las pruebas de simulación AREN

3- SCRIPT_COBERTURA_ENTROPIA.R

El programa contiene las funciones necesarias para correr las pruebas de simulación ARECE

4- PRUEBAS2.R

El programa ejecuta la simulaciones ARE, AREC, ARECS1, ARECS2 y ARECV. Requiere la previa corrida del SCRIPT_2E.R

5-PRUEBAS_CORTE_TURING.R

El programa realiza la simulación ARE con corte. Requiere la corrida previa del SCRIPT_2E.R.

6- PRUEBAS_CORTE_COBERTURA.R

El programa realiza las pruebas de simulación AREC con corte previa corrida del SCRIPT_2E.R

7- PRUEBAS_TURING_ENTROPÍA.R

El programa realiza las pruebas de la simulación AREN. Antes de él debe correrse el SCRIPT_TURING_ENTROPÍA.R

8- PRUEBAS_COBERTURA_ENTROPÍA.R

El programa realiza las pruebas de la simulación ARECE. Antes de él debe correrse el SCRIPT_COBERTURA_ENTROPIA.R

B- ARTÍCULOS PRESENTADOS EN CONGRESOS

1- Santa María, Cristóbal y Soria, Marcelo. “Minería de Datos sobre Comunidades Biológicas”. WICC 2010. El Calafate. Santa Cruz. Argentina

2- Santa María, Cristóbal y Soria, Marcelo. “Aplicaciones de Data Mining al Estudio de la Biodiversidad”. WICC 2011. Rosario. Santa Fé. Argentina.

3- Santa María, Cristóbal y Soria, Marcelo. “Estimación de Biodiversidad por Data Mining y Simulación”. CACIC 2011. La Plata. Buenos Aires. Argentina.